Big Social Data Analytics in Football: Predicting Spectators and TV Ratings from Facebook Data

Nicolai H. Egebjerg[1], Niklas Hedegaard[1], Gerda Kuum[1], Raghava Rao Mukkamala[1] and Ravi Vatrapu[1,2]
[1]Centre for Business Data Analytics, Copenhagen Business School, Denmark
[2]Westerdals Oslo School of Arts, Comm & Tech, Norway
{rrm.itm, rv.itm}@cbs.dk

***Abstract*—This paper explores the predictive power of big social data in regards to football fans' off-line and on-line behaviours. We address the research question to what extent can big social data from Facebook predict the number of spectators and TV ratings in the case of Danish National Football Association (DBU). The predictive model was built from Facebook, match attendance, and TV ratings data sets from 2014-2016. The best fit was a linear regression model with GLM coding. Ultimately, the model did best when predicting the number of spectators based on the Facebook activity during a match as well as the activity from the last two weeks leading up to the match. Furthermore, the data reveals that photos generates the most activity on the national team's page and with videos running at higher production costs there might be some unexploited potential for DBU to improve its social media marketing strategy. Although data limitations are present, this research concludes that predictive models based on big social data can indeed offer important insights for companies to understand their customer base and how to improve marketing strategies.**

***Index Terms*—Big data, Big social media data, Danish National Team, DBU, Facebook data, Football fans, Spectators, TV ratings.**

I. Introduction

A football match is quite an emotional event and being a football team fan evokes a shared sense of emotional attachment to the club, city, and/or country [1]. Fans exhibit social and cultural attachment to clubs[1]. Dansk Boldspil Union (The Danish Football Association) was formed in 1889 with a purpose to promote ball sports, primarily cricket. The Association has, however, since then shifted its focus from other ball sports to primarily focus on football. The organization initially consisted of 86 clubs, including around 4,000 playing members, but has since grown to represent 1,653 clubs and 335,459 members. DBU has thus been one of the main forces in making football the most popular sports in Denmark.The way the association promotes football on a national level is by having the Danish Men's National Football Team play matches against other national teams. These matches can be differentiated into three categories: World Cup qualifiers, European Championship qualifiers, and friendly matches. The Danish public has the possibility to purchase tickets to watch those matches at the stadium, or see the matches live at television. Matches since 2005 have been broadcast on different national media channels, including Kanal 5, 6'eren, Kanal 9, TV 2, TV3, TV3 Puls, and TV3+.

[1]The Social And Community Value Of Football

A. Problem Formulation and Research Question

Since 2010 the Danish National Men's Football Team has faced serious branding issues. Its popularity among Danish citizens has declined and ticket sales have decreased throughout the last 6 years[2]. One of the main reasons for the low popularity has been the football team's unprofessional usage of online and traditional media as a means to create a socio-cultural connection between the team and its fans. As a result, in 2014-16 the association went through radical management changes, the previous men's team head coach Morten Olsen was replaced by ge Hareide, and Claus Bretton-Meyer was announced to become a CEO of DBU. The new CEO decided to solve the low-popularity issue by changing the organizational structure of the association and redefining the national team's brand values[3]. The new CEO of the DBU, Claus Bretton-Meyer (2016) argues that the Danish fans have fallen asleep to a point where only 16% of the Danish population consider themselves either a 'big fan' or a 'very big fan' of the National Team. In 2014, following the CEO succession DBU changed its marketing strategy by creating a new slogan: *A Part of Something Bigger*.In addition, new initiatives were created to increase the football teams (Men, Women, Under-21) presence on online and traditional media. Our paper seeks to explore the efficacy of these new initiatives with regard to social media on spectators and TV ratings. Towards this end, this paper addresses the general research question using the specific case of DBU:

> *To what extent can big social data from Facebook predict fan engagement in terms of spectators and TV Ratings?*

The remainder of the paper is organized as follows. Section 2 explains the conceptual framework. Section 3 presents related work and discusses relevant theories. Section 4 provides a detailed description of the dataset and provides an overview of the process and methods adopted for empirical analysis. Section 5 presents the core empirical findings from the DBU case. Section 6 provides an answer to the research question and a discussion on limitations and implications for future research and practice. Finally, Section 7 provides a short conclusion.

[2]National Team has lost more than a third of its television viewership
[3]Dansk fodbold er en del af noget strre. Berlingske.

II. Related Literature and Detailed Hypotheses

In order to generate demand and produce fan interest, sports leagues justify a range of restrictions that resemble cartels. Szymanski [2, p. 1153] argues that the justification for restrictions can be reduced to three core claims: 1) Inequality of resources leads to unequal competition, 2) fan interest declines when outcomes become less uncertain, and 3) specific redistribution mechanisms produce more outcome uncertainty.The second proposition is of particular interest to this paper. There has been substantial research work in the direction of predicting game attendance. Rottenberg [3] looked at American baseball and argued that *uncertainty of outcome is necessary if the fan/consumer is to be willing to pay admission to the game* (p. 246). Schreyer, Schmidt and Torgler [4] explored the role of Game Outcome Uncertainty (GOU) in season ticket holders' stadium attendance demand and found a positive relationship. Szymanski [2] summarizes research in this area and argues that there seems to be an emerging consensus that demand for match tickets is highest when the home team's probability of winning is about twice that of the visiting team, i.e., a probability of around 0.66 [5] [6].

However, Buraimo and Simmons [7] used TV viewing figures to show that uncertainty of outcome does not have a positive effect on television audience demand. Instead they argue that "there has been a transition of preference for uncertainty of outcome towards a preference for increased talent" [7, p.466]. What attracts spectators and TV viewers is then sporting entertainment performed by superstars. This paper will not dispute that GOU has an effect on the number of spectators. However, it can be mediated by the importance of the particular match. If uncertainty is said to produce interesting matches then it can be argued that matches where the stakes are low (i.e. friendlies) will have less interest and fewer followers. The issue of whether the type of match has an impact on the number of spectators and TV viewers leads us to our first hypothesis:

H1: Matches with high importance (qualifiers) will result in higher TV ratings and number of spectators than matches with low importance (friendlies).

Other related research focuses on the relationship between broadcasting and attendance. Forrest, Simmons and Szymanski [8] studied the English Premier League, which is a cartel of soccer teams that collectively sells the rights to broadcast its matches. Despite considerable demand, the clubs agreed to sell only a fraction of the broadcast rights (60 out of 380 matches played each season between 1992 and 2001). The clubs argued that increased broadcasting would reduce the number of spectators at matches and therefore reduce cartel income. However, the authors found that broadcasting had "a negligible effect on attendance and that additional broadcast fees would be likely to exceed any plausible opportunity cost" [8, p. 243]. If there is a positive correlation between the number of spectators and TV ratings for this data sample it becomes possible to use the two variables interchangeably when answering the research question. Other relevant work studied the relationship between TV ratings and Facebook data with regard to events such as sports broadcasting [9] and talk shows [10].

H2: There is a positive correlation between the number of spectators and TV ratings.

A systematic review of predictive analytics with social media data was conducted by [11] . Researchers have already utilized big social data to predict stock market movements [12] [13], announcements of flu outbreaks [14], forecast revenues for movies ([15], [16]) and to predict election outcomes [17]. Lee, Kim and Cha [16] used a generalized Bass Model (GBM) that reflected both daily seasonality and herd behavior to predict the sales patterns of motion pictures. This is also an interesting model for this paper since football matches might also experience daily seasonality (with higher attendance at matches played on weekends and holidays) and herd behavior.

H3: The match played after a match with a positive result will experience herd behavior and thus have higher attendance than a match played after a negative result.

H4: Matches played on weekdays will have fewer spectators and lower TV ratings than matches played on weekends.

The underlying assumption for this research stream is that social media actions such as tweeting, posting, liking, commenting etc. are proxies for consumer's attention to a particular topic/brand/product and that "the shared digital artefact that is persistent can create social influence" [18, p. 1]. Most related research relies on Twitter data instead of Facebook data. The goal of this paper is to predict ticket sales and TV ratings, which can also be understood as event prediction. They found a 53% correlation between social media activity and ticket sales and furthermore that Facebook had the highest correlation to ticket sales (52%) slightly higher than Twitter (38%). There might also be a difference between the different types of posts on social media in general, and Facebook in particular, and the level of activity they generate. Pletikosa and Michahelles [19] found that different post characteristics had effect on the interaction on Facebook. They did not include videos, but found that photos had the greatest level of engagement followed by statuses and links. Since production value and narrative scope is higher for videos it would be fair to assume that videos will produce more social media activity than other content. This leads to the fifth hypothesis:

H5: Videos will generate more activity than pictures, which in turn will generate more activity than status updates, links and events.

Lassen et al. [18] demonstrated how Twitter data could be used to predict iPhones sales. They developed a linear regression model that transformed iPhone tweets into a prediction of the quarterly iPhone sales. They built their analysis on the AIDA (Attention, Interest, Desire and Action) and Hierarchy of Effects models ([20], [21] in order to understand the relationship between users' propensity to tweet and the probability to purchase the product. This paper will follow the same line of argument. Social media activity surrounding

Landsholdet are associated with all four stages of the AIDA model and all six stages of the Hierarchy of Effects model. Drawing on Asur and Huberman [15] as well as Lassen et al. [18], this paper treats social data from Facebook as a proxy for a user's attention towards the object of analysis, which in this case are matches played by the Danish National Team. Facebook activity is not seen as belonging to a particular stage of the AIDA or HoE models. Instead it is treated as "social media manifestations of real-world activities" [18, p. 83] of fans/consumers with respect to football matches. This leads to our sixth and main hypothesis for this paper:

H6: Matches played during periods with high Facebook activity will have more spectators and higher TV ratings than matches played during periods with less Facebook activity.

III. Methodology

A. Dataset Description

Two data sets were used in this paper: Facebook data and match data. The first data set contained data from the Danish National Team's official Facebook page. The raw data consisted of a little more than 2.1M data points where each row is equivalent to an action on the Facebook page. The data contains information on action type (whether it is a post, comment or like), actor name and ID, timestamp, type of post, and if relevant, links and text value for posts and comments. The social data available ranges from 10/30/2014 to 11/10/2016 which covers 11 matches played during that time period. The aggregated Facebook data was ordered in dimensions of total posts, comments and likes for each match over a two-week, one-week and two-days window and during the event as shown in table II.

The second data set about matches contained information such as date of the match, number of spectators, TV viewer ratings and other control variables needed to test the hypotheses such as the result, type of match and the broadcasting channel as shown in table I. The data was collected for all home games played between 2005 and 2016. In order to answer the research question only data from 2014-2016 was necessary, but the additional data provides some interesting insights and is required to test the secondary hypotheses.

B. Data Analysis Process

The data analysis process is illustrated in figure 1. Altogether 2,132,003 data entries from the Facebook page of *Landsholdet* were collected using the tool SODATO [22] [23] and TV ratings were collected from TNS Gallup. The two data sources were then combined using Tableau and SAS Studio - tools that were later used for descriptive-, visual- and predictive analytics in order to answer the research question by hypotheses testing.

[4]WC Q: World Cup qualifier, Euro Q: Euro Cup Qualifier

[5]Country codes KZ:Kasakhstan, ME:Montenegro, AM:Armenia, LI:Liechtenstein, IS:Iceland, SE: Sweden, FR:France, AL: Albania, RS:Serbia

Date	Country Code	Type[4]	Total spectators	delta spectators	Res	TV rating	Week	Channel
2016/11/11	KZ	WC Q	18901	-1681	4-1		45	CH5
2016/10/11	ME	WC Q	20582	-1213	0-1	650	41	CH5
2016/09/04	AM	WC Q	21795	13791	1-0	620	35	CH5
2016/08/31	LI	Friendly	8004	-1190	5-0	302	35	CH5
2016/03/24	IS	Friendly	9194	-26857	2-1	452	12	CH5
2015/11/17	SE	Euro Q	36051	17906	2-2	900	47	CH5
2015/10/11	FR	Friendly	18145	-17503	1-2	305	41	CH5
2015/09/04	AL	Euro Q	35648	4761	0-0	810	36	CH5
2015/06/13	RS	Euro Q	30887	21707	2-0	651	24	CH5
2015/06/08	ME	Friendly	9180	-1325	2-1	328	24	CH5
2015/03/25	USA	Friendly	10505	10505	3-2	458	13	CH5

Table I
INFORMATION ABOUT MATCHES AND SPECTATORS

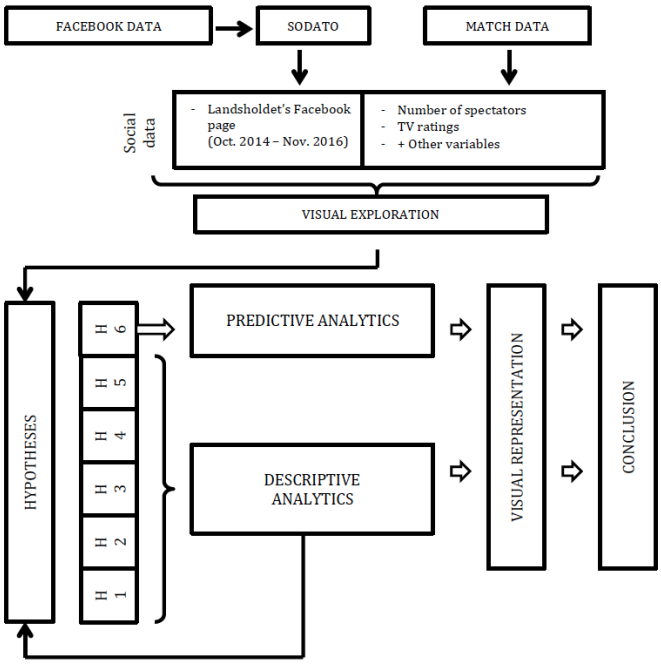


Figure 1. Data Analysis Process Diagram

For the prediction model in H6 different statistical models were evaluated. The final choice was to use a multiple regression model with GLM coding in SAS studio. Control variables were included based on findings from the previous hypotheses (H1 and H4). By including match type and day of the match the correlation was improved resulting in a RMSE of 1.762 for the number of spectators (compared to 8.230 without control variables). Additionally, we tested whether past results could work as a predictor of herd behaviour or a trend. However, when this variable of past results was added to the model it was no longer statistically significant. This is probably an effect of too few observations in the sample. Finally, inputs for the prediction model were:

$$Y = \alpha + \beta_0 * F_w + \beta_1 * F_m + \beta_2 * M_t + \beta_3 * S_d + \epsilon_{wmtd}$$

where F_w: total Facebook activity leading up to the matches,

Countries[5]	KZ	ME	AM	LI	IS	SE	FR	AL	RS	ME	USA	Total
	Nov 16	Oct 16	Sep 16	Aug 16	Mar 16	Nov 15	Oct 15	Sep 15	Jun 15	Jun 15	Mar 15	
2 Weeks Before												
Posts	N/A	256	304	148	124	398	280	220	216	90	172	2,208
Comments	N/A	6,913	1,481	308	765	24,230	5,529	14,133	9,617	2,438	4,933	70,347
Likes	N/A	42,154	80,822	31,414	46,876	147,073	103,893	55,779	100,610	44,116	34,536	687,273
1 Week Before												
Posts	N/A	200	246	104	100	296	224	174	168	66	112	1,690
Comments	N/A	6,739	1,327	141	679	22,625	5,003	13,070	8,113	1,660	4,460	63,817
Likes	N/A	35,158	71,011	23,399	45,218	117,615	92,843	42,192	67,726	36,406	26,246	557,814
2 Days Before												
Posts	N/A	116	76	58	42	84	44	102	54	24	60	660
Comments	N/A	6,363	139	138	404	14,568	789	12,743	6,538	218	3,822	45,722
Likes	N/A	6,921	18,712	16,827	18,726	35,783	12,986	28,957	20,222	7,708	15,080	181,922
3 Hours During The Match												
Posts	N/A	74	38	34	30	102	32	50	40	54	52	506
Comments	N/A	595	72	18	169	1,961	496	1,280	750	236	1,064	6,641
Likes	N/A	2,980	25,487	13,384	20,281	11,098	5,822	8,050	68,064	23,584	35,674	214,424

Table II
FACEBOOK DATA FOR HOME MATCHES

	Spectators			Tv Ratings		
Model	Root MSE	P-Value	R-Square	Root MSE	P-Value	R-Square
2 Week Act	2746.46	0.0008	0.965	113423	0.0017	0.712
1 Week Act	2494.15	0.0005	0.971	113927	0.0018	0.710
2 day Act	2974.37	0.0011	0.959	99278	0.0003	0.780
2 Week + During	1776.01	0.0006	0.989	118041	0.0054	0.712
1 Week + During	1762.82	0.0006	0.988	118575	0.0057	0.710
2 Day During	2715.78	0.0031	0.973	101662	0.001	0.787
During	2567.79	0.0005	0.969	118588	0.0029	0.686

Table III
MODELS FOR PREDICTING SPECTATORS AND TV RATINGS

continuous variable (different windows were used: 2 weeks, 1 week and 2 days)
F_m: total Facebook activity during the match, continuous variable (3.5 hours before, 2 hours during and 0.5 hours after the match)
M_t: match type (categorical variable: WC Qualifier, Euro Qualifier and Friendly)
S_d: daily seasonality based on the day of the match (categorical variable: weekday or weekend)
Y: number of spectators/TV viewers.

The primary coefficients of interest are 0 and 1 which can be interpreted as the contribution of social media activity to the number of spectators or TV viewers that will watch the match. However, due to the introduction of the control variables these coefficients may be negative although they correlate positively with the dependent variable when standing alone. All the different test combinations of the model effects are presented in table III. Based on the RMSE the best model for predicting the number of spectators used Facebook data from 1 weeks leading up to, as well as the activity during the match. Alternatively, in order to predict TV ratings, using Facebook data from 2 days prior to the match proved to be the best fit as shown in table III.

IV. FINDINGS

In this section, we present the results of our data analysis and discuss whether the hypotheses were confirmed or rejected. One of the important findings is that the most

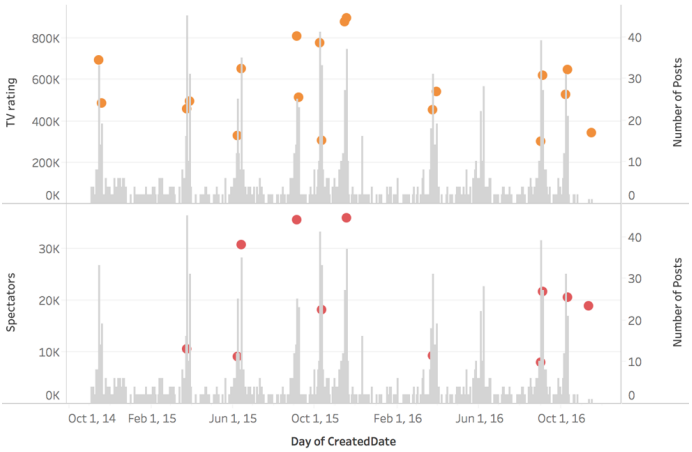
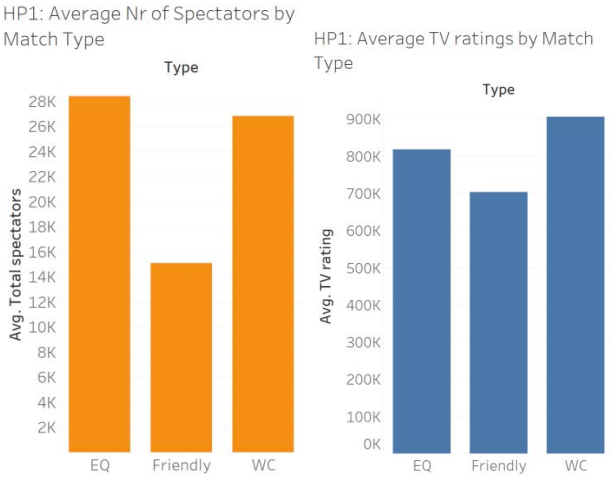


Figure 2. Distribution of Facebook posts vs. spectators and TV ratings

of the activity on their Facebook page is generated around matches. Especially the temporal distribution of various Facebook actions (such as posts, postlikes, comments and so on) indicated large amount of peaks before and during the match events. Moreover, there is also significant visual coherence between the Facebook actions verses number of spectators and TV ratings. Figure 2 shows one such distribution where the distribution of Facebook posts by DBU verses spectators and TV ratings is plotted.

*1) **Hypothesis H1**:* Hypothesis 1 was tested for both the match data gathered from 2005-2016 as well as on the sample data. The visual analytics clearly showed that qualifiers had both higher TV ratings and spectators as shown in figure 3. When tested in SAS, there is significant correlation on both TV



(a) Spectators (b) TV Ratings

Figure 3. Spectators and TV-ratings for each match type

ratings and number of spectators for the entirety of the match data with P-values of 0.0126 and 0.0001 respectively. Here we only distinguished between qualifiers and friendlies. In the sample data, the types of qualifier was distinguished from each other. However, the correlation still holds with Euro qualifiers compared to friendlies having P-values of 0.0001 for spectators and 0.0001 for TV ratings. World Cup (WC) qualifiers show the same pattern with P-values of 0.0113 for spectators. The correlation between WC qualifiers and friendlies has a P-value is 0.0525. However, seen together with the entire match data this would be statistically significant. Therefore, **H1** is confirmed, and matches with higher importance will have higher TV ratings and number of spectators.

*2) **Hypothesis H2**:* A correlation analysis of the number of TV viewers and number of spectators was done in SAS Studio. A visual representation of this can be seen in figure 4. It is difficult to see with the naked eye whether there is a correlation here or not, so here a calculation was needed. The analysis returned *Total spectators = 0.0170159 * TV rating + 6486.7* at a significance level of $p < 0.0001$ and therefore there exists a clear correlation between the two. Thus **H2** is confirmed. This means that when the numbers of spectators are growing, so is the number of people watching on TV and vice versa.

*3) **Hypothesis H3**:* In order to calculate whether matches played after a positive result experienced herd behaviour the authors had to calculate a *delta spectators* i.e. the change in spectators from match to match and a fixed result. The fixed result is calculated by taking the difference in goals, e.g. a 3-1 defeat is calculated as a -2. As shown in figure 5 there seems to be an outlier in the upper left corner. This match where the fixed result is -4 resulted in an increase in spectators for the next home game of more than 20,000. The correlation between the two is however still significant with $p = 0.049 < 0.05$ and with an plot equation $DeltaSpectators(noFriendly) = 1896.79 * Result fixed - 1963.86$. When the outlier is removed the correlation becomes stronger with $p = 0.0006 < 0.01$.

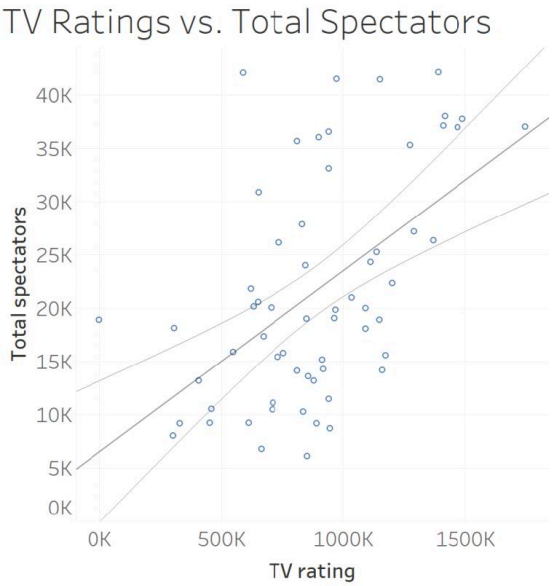


Figure 4. TV ratings verse spectators

Thus **H3** is confirmed. This means that the game played before a home match will have an impact on the number of spectators for the next home game

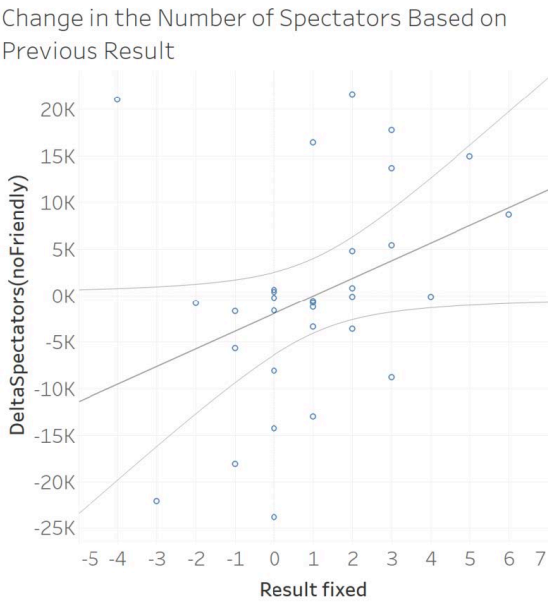


Figure 5. Delta spectators vs previous match's result

*4) **Hypothesis H4**:* The overall day-wise distribution of TV-ratings and total spectators is shown in figure 6. In order to investigate H4 the days of the week was binary coded. Monday through Thursday was coded as 0 for weekday, and Friday to Sunday as 1 for weekend. Tableau was used to analyze visually whether H4 was true. The weekend matches have a much higher number of spectators 25,102 vs. 19,777. However, weekday matches seems to have a higher number of viewers on TV with 901,725 vs. 865,048 on average. This

suggests that either people watch something else during the weekend or use their time on other things than watching TV. Thus **H4** is only partially confirmed.

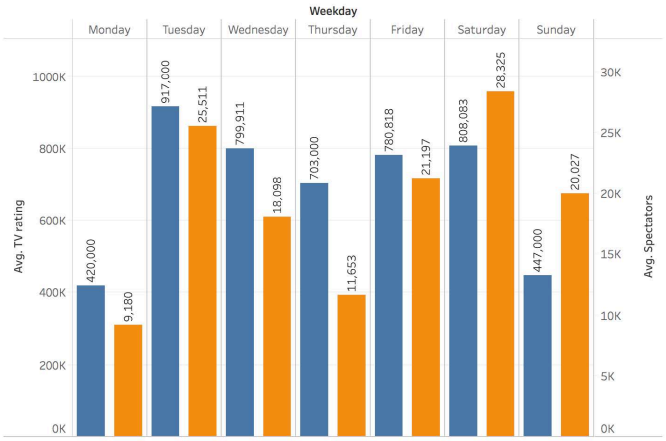


Figure 6. Day wise distribution of TV ratings and total spectators

*5) **Hypothesis H5**:* H5 was tested by taking all activity on the different kind of posts and then comparing the average activity as shown in figure 7. It shows that the post types experiencing the most activity are photos followed by statuses and videos. Our result is in accordance with the findings of [24] which showed that *photo* has high engagement potential among all post types of Facebook. We further analyzed to see the correlation between the different post types and it shows that photos will always receive more activity on average than all other post types except for statuses. There were no other correlations between the posts types. Thus **H5** is only partially confirmed. The results of **H5** suggest that consumers don't take the time to watch videos on Facebook. The extra time and cost it takes for Landsholdet to produce videos is thus not worthwhile and it is suggested that they decrease the number of videos on their Facebook wall. In any case this result suggests that it could be useful to change the way that Landsholdet does videos on Facebook. That is, they might have to change the content of the videos or the length of them. At the same time since photos are vastly superior compared to other posts types it is suggested that they increase the number of posts on their Facebook wall in order to create extra activity. The suggestions here raise a couple of questions. When is enough? When will photos stop creating extra activity, and are they only superior at the moment because they enter into a mix of different post types? The overall marketing strategy was not studied in this paper - and the mix of post types that Landsholdet uses might be a deliberate move in their branding activities.

*6) **Hypothesis H6**:* The hypothesis **H6** is split into two sub hypothesis (H6a and H6b) to predict the number of spectators and the TV ratings respectively. As shown in table III, different models were tested in order to find the ones most accurately predicting the number of spectators and the TV ratings. The most accurate model for spectators was the one with all the Facebook activity from 1 week leading up to and the activity

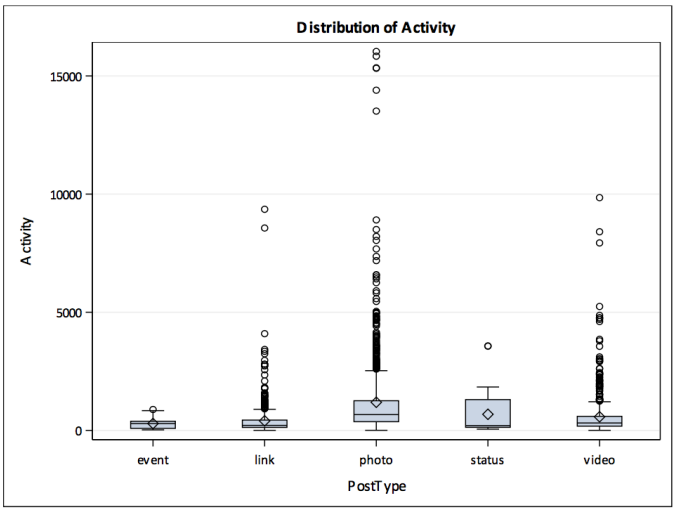


Figure 7. Day wise distribution of TV ratings and total spectators

during a match. The one best predicting TV ratings included Facebook activity from the two days before a match. In both cases increased Facebook activity has a positive effect on the dependent variables. Thus both H6a and H6b are confirmed.

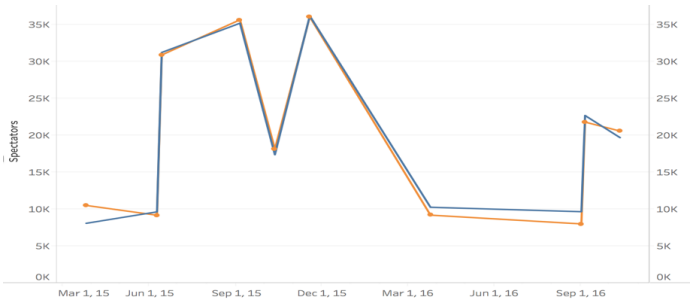


Figure 8. Predictive model for spectators with 1 week + during forecasting model

The results of the predictive model of total spectators and TV ratings can be seen in figure 8 and 9 respectively, where the red lines indicates predicted values and dark blue lines indicate actual values. Moreover, the multiple linear regression model results for spectators and TV ratings are presented in table IV. One could notice that the multiple linear regression model results with a high value of adjusted R-square (≈ 0.97) indicates good amount of fit as also indicated in figure 8. For TV ratings, the model results are reasonably satisfactory (adjusted R-square ≈ 0.71) with a fair amount fitness as can also be seen in figure 9.

V. Discussion

This paper investigated the consequences of the specific case of DBU's new digital media strategy in terms of total number of spectators and TV ratings based on user engagement on DBU's official Facebook wall.

First, we found that DBU can improve their digital media strategy by making fewer video posts on Facebook and instead post more photos as they carry more engagement potential than

Spectators		TV Ratings	
Root MSE	1762.81627	**Root MSE**	99278
Dependent Mean	19999	**Dependent Mean**	577444
R-Square	0.9887	**R-Square**	0.78
Adj R-Sq	0.9745	**Adj R-Sq**	0.7123
AIC	164.33045	**AIC**	438.34699
AICC	220.33045	**AICC**	445.98335
SBC	154.14596	**SBC**	422.79885

Table IV
MULTIPLE LINEAR REGRESSION MODEL RESULTS FOR SPECTATORS AND TV RATINGS

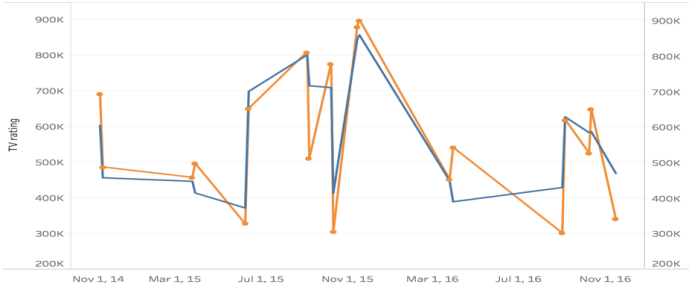


Figure 9. Predictive model for TV ratings with 2-day forecasting model

videos. Our finding is contrary to conventional wisdom that posits that as football is an active field, game videos must be more appealing to the people than photos. That said, our finding also confirms the [24] that the Facebook post type of photo caries high engagement potential .

Second, we also found that social media data is indeed able to predict the number of spectators and the TV ratings of football matches fairly accurately. Unlike previous work, this was done using neither the Game Outcome Uncertainty (GOU) nor the Quality of the players. This suggests that GOU, as Buraimo & Simmons [7] observed, is not the only variable affecting spectator attendance and TV ratings. However, both the variables mentioned here could very likely strengthen the predictive models of this paper. There are many variables influencing demands for football tickets and a handful of them were included in this study. A few that were not included here are weather, GOU and Star Quality of the players. In addition, it would have been useful to include business data showing continuous sales and information about season ticket holders. This would have allowed for more accurate analytics of how Facebook data influence sales of companies. However, this would also raise ethical issues as to how closely people purchasing tickets and season ticket holders should be monitored. The models applied here only used the total activity without ever including information about actors on the social medium. Thus one could argue that the privacy of the individual is more secure in this version. Third, until now only limited research has focused on whether Facebook social data can predict sales patterns with previous research mostly focusing on Twitter data. Future research would have to look into other areas. However, as it stands, it seems likely that companies can influence their own sales by posting content about their products on social media.

Fourth, this study could have included textual analysis in order to investigate the sentiment towards Landsholdet. This would have provided a more precise indication of the mood of the posts and given additional information for predicting the number of spectators. This would also in some cases have indicated who actually attends or watches the games from their sofa. However, once more the ethical issue of how closely these individuals should be monitored resurfaces. In any case, future research into the predictive capabilities of Facebook data would benefit from including some sort of textual analysis.

Overall, the findings in this paper indicate that football clubs should increase their social media presence and make sure that they post content on a continuous basis since it creates demand for tickets irrespective of how likely an outcome is.

A. *Recommendations for Case Company*

The outcome of this study indicates that DBU at the moment does not utilize big data in their marketing strategy. If this was the case they would have known the diminished return on the time invested in creating videos. Using big data could also help them recognize which actors are the biggest fans and thus aid them in their communication towards those. However, it might be that at present they do not have the resources to do so.

In the short-term DBU should investigate the connection between fan activity on their Facebook page and ticket sales. Analyzing continuous sales and social media activity together might provide them with even better tools to understand what type of posts and content that drives sales and fan interest. In the mid- to long-term DBU should continue to work on their brand image. It is now clear that social media is, and should be, part of their marketing strategy. Generating content at the right time, targeted towards the right fan base will eventually help them to increase the amount of loyal fans. As the sample data illustrates, most of the activity on their Facebook page is generated around matches as shown in figure 2. Since the national team only plays 5-6 matches a year it becomes necessary to focus on the season breaks and silent periods between matches.

However, there might be diminishing returns to sharing content online, which is something to be cautious about. An organisation like DBU must be careful not to create posts that could be understood as *clickbait* since the goal of these sites is often high traffic and low engagement[6], while selling tickets requires high engagement. Instead, DBU should follow other companies that use machine learning and sophisticated recommendation algorithms that identify potential customers and send them messages such as *other fans of Danish football bought tickets for this game* at key points along the decision journey. A study by McKinsey found that these algorithms are highly effective at converting customers, though with an important limitation: *the influence ... can be as much as 75 percent lower if messages aren't highly personalized and targeted* [25].

[6]The dirty secrets of clickbait. This post will blow your mind!

B. Limitations of the Study

This paper has three limitations. First, social data was only available for little more than two years. Having data for more years could have made it possible to see the effects of the new marketing strategy launched after the appointment of Bretton-Meyer as CEO. Second, the data on ongoing sales or season ticket holders is not available. The latter has been the primary focus of DBU for the past two years. Third and last, the TV ratings predictive model could have been improved had there been a control variable for the pull towards other channels during match time.

VI. Conclusion

By using data fetched from the Danish national football team's Facebook page it was possible to set up a predictive model for the number of spectators and TV viewers. It is a fairly simple model relying on only two other control variables: match type and day of the match. However, since there were various data and resource limitations, the models could be improved even further. These limitations include the fact that very few matches were played during the sample period, no distinction was made between positive/neutral/negative posts, no data was available for ongoing sales or season ticket holders and not taking other channels into consideration for the TV ratings-model.

Assuming increased activity leads to more spectators and higher TV ratings (the sample shows mixed results), DBU can improve upon their social media marketing strategy by making better returns on their video posts. Although production costs for videos are higher, it is currently their photo posts that generate the most activity among fans. Furthermore, by posting improved content more often, also between matches, while avoiding clickbait, they should see an increase in season ticket holders, which is their primary concern for the future. Future work should gather more data and do sentiment analysis to see how this would affect the predictive model.

VII. Acknowledgements

We thank members of the Centre for Business Data Analytics (bda.cbs.dk) for their feedback on the paper. The authors were partially supported by the project Big Data Analytics for Social Business funded by the Industriens Fond (The Danish Industry Foundation). Any opinions, findings, interpretations, conclusions or recommendations expressed in this paper are those of its authors and do not represent the views of the Industriens Fond.

References

[1] I. Abosag, S. Roper, and D. Hind, "Examining the relationship between brand emotion and brand extension among supporters of professional football clubs," *European Journal of marketing*, vol. 46, no. 9, pp. 1233–1251, 2012. 1

[2] S. Szymanski, "The economic design of sporting contests," *Journal of economic literature*, vol. 41, no. 4, pp. 1137–1187, 2003. 2

[3] S. Rottenberg, "The baseball players' labor market," *Journal of political economy*, vol. 64, no. 3, pp. 242–258, 1956. 2

[4] D. Schreyer, S. L. Schmidt, and B. Torgler, "Against all odds? exploring the role of game outcome uncertainty in season ticket holders' stadium attendance demand," *Journal of Economic Psychology*, vol. 56, pp. 192–217, 2016. 2

[5] G. Knowles, K. Sherony, and M. Haupert, "The demand for major league baseball: A test of the uncertainty of outcome hypothesis," *The American Economist*, vol. 36, no. 2, pp. 72–80, 1992. 2

[6] D. Forrest and R. Simmons, "Outcome uncertainty and attendance demand in sport: the case of english soccer," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 51, no. 2, pp. 229–241, 2002. 2

[7] B. Buraimo and R. Simmons, "Uncertainty of outcome or star quality? television audience demand for english premier league football," *International Journal of the Economics of Business*, vol. 22, no. 3, pp. 449–469, 2015. 2, 7

[8] D. Forrest, R. Simmons, and S. Szymanski, "Broadcasting, attendance and the inefficiency of cartels," *Review of Industrial Organization*, vol. 24, no. 3, pp. 243–265, 2004. 2

[9] A. Hennig, A.-S. Åmodt, H. Hernes, H. M. Nygårdsmoen, P. A. Larsen, R. R. Mukkamala, B. Flesch, A. Hussain, and R. Vatrapu, "Big social data analytics of changes in consumer behaviour and opinion of a tv broadcaster," in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016, pp. 3839–3848. 2

[10] H. H. Larsen, J. M. Forsberg, S. V. Hemstad, R. R. Mukkamala, A. Hussain, and R. Vatrapu, "Tv ratings vs. social media engagement: Big social data analytics of the scandinavian tv talk show skavlan," in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016, pp. 3849–3858. 2

[11] N. B. Lassen, L. la Cour, and R. Vatrapu, "Predictive analytics with social media data," *The SAGE Handbook of Social Media Research Methods*, p. 328, 2017. 2

[12] Y. Karabulut, "Can facebook predict stock market activity?" *AFA 2013 San Diego Meetings*, 2013. 2

[13] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of computational science*, vol. 2, no. 1, pp. 1–8, 2011. 2

[14] V. Lampos and N. Cristianini, "Tracking the flu pandemic by monitoring the social web," in *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*. IEEE, 2010, pp. 411–416. 2

[15] S. Asur and B. Huberman, "Predicting the future with social media," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, vol. 1, 2010, pp. 492–499. 2, 3

[16] Y. Lee, S.-H. Kim, and K. C. Cha, "A generalized bass model for predicting the sales patterns of motion pictures having seasonality and herd behavior," *Journal of Global Scholars of Marketing Science*, vol. 22, no. 4, pp. 310–326, 2012. 2

[17] D. Gayo-Avello, "A meta-analysis of state-of-the-art electoral prediction from twitter data," *Soc. Sci. Comput. Rev.*, vol. 31, no. 6, pp. 649–679, Dec. 2013. 2

[18] N. B. Lassen, R. Madsen, and R. Vatrapu, "Predicting iphone sales from iphone tweets," in *Enterprise Distributed Object Computing Conference (EDOC), 2014 IEEE 18th International*. IEEE, 2014, pp. 81–90. 2, 3

[19] I. P. Cvijikj and F. Michahelles, "A case study of the effects of moderator posts within a facebook brand page," in *International Conference on Social Informatics*. Springer, 2011, pp. 161–170. 2

[20] H. Li and J. D. Leckenby, "Examining the effectiveness of internet advertising formats," *Internet advertising : theory and research*, pp. 203–224, 2007. 2

[21] R. J. Lavidge and G. A. Steiner, "A model for predictive measurements of advertising effectiveness," *Journal of Marketing*, vol. 25, no. 6, pp. 59–62, 1961. [Online]. Available: http://www.jstor.org/stable/1248516 2

[22] A. Hussain, R. Vatrapu, D. Hardt, and Z. Jaffari, "Social data analytics tool: A demonstrative case study of methodology and software." in *Analysing Social Media Data and Web Networks*, M. C. Rachel Gibson and S. Ward, Eds. Palgrave Macmillan, 2014 (in press). 3

[23] R. Vatrapu, R. R. Mukkamala, A. Hussain, and B. Flesch, "Social set analysis: A set theoretical approach to big data analytics," *IEEE Access*, vol. 4, pp. 2542–2571, 2016. 3

[24] N. Straton, K. Hansen, R. R. Mukkamala, A. Hussain, T.-M. Gronli, H. Langberg, and R. Vatrapu, "Big social data analytics for public health: Facebook engagement and performance," in *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 2016, pp. 1–6. 6, 7

[25] J. Bughin, "Gettig a sharper picture of social media?s influence," *McKinsey Quartely, July*, 2015. 7