Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

The added value of auxiliary data in sentiment analysis of Facebook posts

Matthijs Meire[a], Michel Ballings[b,*], Dirk Van den Poel[a]

[a] Department of Marketing, Ghent University, Tweekerkenstraat 2, Ghent, 9000 Belgium
[b] Department of Business Analytics and Statistics, The University of Tennessee, 249 Stokely Management Center, 916 Volunteer Blvd, Knoxville, 37996 TN, USA

ARTICLE INFO

Article history:
Received 10 September 2015
Received in revised form 17 June 2016
Accepted 19 June 2016
Available online xxxx

Keywords:
Facebook
Text mining
Sentiment analysis
Machine learning
Social media

ABSTRACT

The purpose of this study is to (1) assess the added value of information available before (i.e., leading) and after (i.e., lagging) the focal post's creation time in sentiment analysis of Facebook posts, (2) determine which predictors are most important, and (3) investigate the relationship between top predictors and sentiment. We build a sentiment prediction model, including leading information, lagging information, and traditional post variables. We benchmark Random Forest and Support Vector Machines using five times twofold cross-validation. The results indicate that both leading and lagging information increase the model's predictive performance. The most important predictors include the number of uppercase letters, the number of likes and the number of negative comments. A higher number of uppercase letters and likes increases the likelihood of a positive post, while a higher number of comments increases the likelihood of a negative post. The main contribution of this study is that it is the first to assess the added value of leading and lagging information in the context of sentiment analysis.

1. Introduction

In the beginning of the century, Web 2.0 emerged as an ideological and technical foundation giving rise to the massive production of user generated-content (UGC). Blogging platforms and online retailers are the first examples of this foundation [50]. Today, UGC is still growing rapidly, sparking interest and activity in opinion mining and sentiment analysis [62, 74]. Sentiment analysis is defined as the computational process of extracting sentiment from text [61, 74]. Applications range from the prediction of election outcomes [17, 92], to relating public mood to socio-economic variables [17], to improved e-learning strategies [72].

Early examples of sentiment analysis were mainly based on review data. This type of data rarely contained much more information than the content and the time of posting of the review itself. Models using these data are based on present information, where 'present' refers to the time of posting. This changed with the advent of social networks such as Facebook and Twitter in that much more data became available. On these platforms, not only the focal post's content is available, but, taking into account the time of posting, there is also leading and lagging information. Leading information is available even before content is posted (e.g., user profiles, previous posts) and thus contains information about the past. On the other hand, lagging information is generated a posteriori, after the content was posted (e.g., interactions such as likes or retweets) and thus contains information about the future (seen from the time of posting). Leading information can therefore be included in any sentiment model, while lagging information can be included in tools that do not require real-time sentiment analysis. To the best of our knowledge, there is no study that includes leading and lagging information into sentiment analysis models. However, we believe that we can improve sentiment prediction by including leading and lagging information for several reasons. First, social media suffer from a lot of slang [41, 72] making it harder for traditional methods to achieve satisfactory model performance on text variables alone. Second, leading variables would take into account users' average sentiment, word use, well-being, and mood and demographics, effectively acting as a user-specific informative prior of future sentiment and accounting for heterogeneity among users. Leading variables have been shown to lead to better predictions [10]. Third, extant literature has found significant relationships between post sentiment and lagging information such as likes and comments [87].

To fill this gap in literature, we assess the additional value for sentiment analysis of leading and lagging information over and above information extracted from the focal post. We do this by constructing three models. The first model is the base model that focuses on the present and contains only the focal post (including text and timing of posting). The second model contains both the focal post's content and leading information, and thus contains both present and past

[*] Corresponding author.
E-mail addresses: Matthijs.Meire@UGent.be (M. Meire), Michel.Ballings@utk.edu (M. Ballings), Dirk.VandenPoel@UGent.be (D. Van den Poel).

http://dx.doi.org/10.1016/j.dss.2016.06.013
0167-9236/© 2016 Elsevier B.V. All rights reserved.

information. Finally, the third model augments the second model with lagging information. This means that the third model takes into account the past, present and future information of a post.

The remainder of this article is structured as follows. First, we provide a literature review focusing on sentiment analysis of social media data and the reasons why leading and lagging information might be valuable in a sentiment prediction model. Second, we detail our methodology including the data, the model description, the predictors, the predictive algorithms and the model evaluation measure. The third section discusses the results. The penultimate section consists of the conclusion and practical implications of this research. In the final section we address the limitations and avenues for future research.

2. Literature review

There are two main approaches to sentiment analysis [72, 88]. The first approach consists of lexicon-based models, which use predefined lexicons of positive, neutral and negative words to assign positivity values to a sentence or text (e.g., [46, 93]). Machine learning-based methods constitute the second approach. These methods use several text features (e.g., syntactic features and lexical features; we refer to McInnes [64] for a complete overview of these features) as input for a training model and predict the sentiment of text using these features [88]. Machine learning methods have been shown to be more accurate than lexicon-based methods in general, but also more time consuming [20, 75]. Lexicon-based methods, however, tend to perform better in less-bounded domains [72]. Recently, the two approaches have been combined by several authors [58, 65, 72, 90, 98], mostly by using the scores from a lexicon-based exercise as input features for the machine learning algorithm. In this study we will adopt such a hybrid approach. The reason is that the approach allows for additional features to be added to the model.

Literature on sentiment analysis can be summarized according to (1) the use of a focal post's features [64] , (2) the use of auxiliary features [10] , and (3) the focal post's source [1] . The focal post's features constitute: (1) lexicon features, which denote either a pure lexicon-based approach or a combination of lexicon and machine learning, (2) lexical features (bag-of-words, n-grams, co-occurrence and collocations), (3) syntactic features (morphology, part-of-speech) and (4) time features. The auxiliary features are divided into leading and lagging features. The former denotes all the information, with regard to a specific user, that is available until the moment of posting. The latter includes information that is available one week after posting (i.e., information on the likes and the comments a post has received). Stated differently, the focal post's features reflect all information of the present, where 'the present' refers to the time of posting, which will be different for every post. Every action that occurred before the present, is referred to as 'the past', while 'the future' indicates all actions that occurred after posting. The leading variables thus originate in the past, while the lagging variables originate in the future.

Table 1 provides a representative overview of literature with a focus on social media applications, as social media contain leading and lagging information. It is apparent that sentiment analysis has been widely applied to a diverse set of social media. Table 1 shows that both the lexicon-based (denoted an x in the column labeled 'Lexicon') and the machine learning approaches have been used, and that plenty of text features have been explored. However, it also shows that there is a large potential source of information for sentiment analysis that remains largely untapped. Indeed, social media do not

Table 1
Literature overview.

	Features of focal post				Auxiliary features		Text source
	Lexicon	Lexical	Syntactic	Time	Leading	Lagging	
Pang et al. [75]		x	x				Reviews
Dave et al. [26]		x	x				Reviews
Yu and Hatzivassiloglou [96]	x	x	x				News Items
Bai et al. [4]		x					Reviews
Gamon [40]		x	x				Customer feedback
Mullen and Collier [68]		x					Reviews
Matsumoto et al. [63]		x					Reviews
Read [80]		x					Reviews
Riloff et al. [81]		x	x				Reviews
Abbasi et al. [1]		x	x				Reviews
Go et al. [41]		x	x				Twitter
Prabowo and Thelwall [78]	x	x	x				Reviews
Melville et al. [65]	x						Reviews
Pak and Paroubek [73]		x					Twitter
Barbosa and Feng [9]	x		x				Twitter
Davidov et al. [27]		x					Twitter
Kouloumpis et al. [53]	x	x	x				Twitter
Taboada et al. [88]	x						Reviews
Agarwal et al. [2]	x	x	x				Twitter
Smeureanu and Bucur [85]		x					Reviews
Wang and Manning [94]		x					Reviews
Neri et al. [69]		x					Facebook
Blamey et al. [15]		x					Twitter
Kumar and Sebastian [56]	x	x					Twitter
Ben Hamouda and El Akaichi [13]		x					Facebook
Troussas et al. [91]		x					Facebook
Tamilselvi and ParveenTaj [89]		x	x				Twitter
Habernal et al. [42]		x	x				Facebook
Ortigosa et al. [72]	x						Facebook
Basiri et al. [10]	x	x			x		Reviews
da Silva et al. [24]		x					Twitter
Fersini et al. [36]		x					Reviews, Twitter
Yu and Wang [97]		x					Twitter
Mohammad and Kiritchenko [67]		x					Twitter
Our study	x	x	x	x	x	x	Facebook

only offer an efficient way to gather the focal post's textual data used in traditional sentiment analysis, they also allow to gather a lot of auxiliary data (e.g., user profile information, likes on statuses) that have not yet been used in sentiment analysis. Basiri et al. [10] recently made an effort to incorporate such data into a sentiment analysis model. They found that deviations of a reviewer's post compared to the previous posts of this same reviewer lead to better review score prediction. The model of Basiri et al. [10] is, however, limited to the incorporation of one auxiliary variable and therefore does not reflect the full potential. Furthermore, they do not incorporate the leading information into a sentiment analysis model, but only use it for the prediction of review scores.

In this study we will exploit the focal post's information as well as auxiliary leading and lagging data that are present on Facebook. This allows us to assess the improvement in the prediction of emotional valence of Facebook statuses that stems from incorporating auxiliary data. The following section clarifies why leading and lagging information may be important (i.e., improve the predictive performance of our models). This information is also summarized in the conceptual framework depicted in Fig. 1.

2.1. Leading information

Leading Facebook information includes the complete history of a user's Facebook trail, including previous posts. We hypothesize that this information will improve sentiment classification prediction because several user characteristics can influence expressed sentiment. Settanni et al. [84] show that textual indicators extracted from Facebook may be used to study subjective well-being, a result confirmed by Kramer [55]. This means that, by looking at previous posts of the same user and the valence of those posts, we can make an assumption about the subjective well-being of the user. Moreover, Diener [32] states that personality is a major determinant of long-term, subjective well-being. This is an important point, given that several researchers report that Facebook profile features [51, 71] as well as text [76] can accurately predict personality traits. By incorporating these Facebook profile features and previous textual features, we thus aim to incorporate the subjective well-being of a user as a predictor. As this is a long-term emotional state of a user, we believe subjective well-being can be informative of the sentiment expressed in Facebook posts.

While subjective well-being can add value, short-term changes (the 'mood' of a user) can affect sentiment of Facebook posts as well. Ortigosa et al. [72] state that behavior variations as shown on Facebook, can indicate changes in the user's mood. Smith and Petty [86] report that positive or negative framing of a message could create more attention, especially in the case where the framing is unsuspected, as is the case with short-term changes from subjective well-being. We therefore argue that deviations from a user's average posting behavior can be informative of the sentiment of that post.

Comparable to a person's subjective well-being, we refer to network well-being as the overall emotional state of the network of the user. Network well-being and the focal user's well-being are connected by a phenomenon called emotional contagion [21, 47], which is defined as the tendency to automatically mimic other persons, and consequently to converge emotionally [45]. This influence works in both ways. Network well-being can thus be informative about a user's well-being, and hence about the sentiment expressed in the user's Facebook posts. Quercia et al. [79] already showed that community well-being can be predicted by using sentiment of community members' tweets. Since Facebook posts of the user's network were not available, we use the reactions to previous posts of the focal user to take into account part of network well-being that can be measured.

Finally, Schwartz et al. [83] not only found differences in language usage across personalities, but also across gender and age. By incorporating these demographic variables and allowing for interaction effects, we assume that the textual features can bring even more added value to sentiment prediction.

Overall, the leading variables allow researchers to take into account heterogeneity among users with regards to word use, well-being, mood and demographics. The leading variables are discussed in detail in Section 3.4.2 and Table A2 in Appendix A. Fig. 1 shows the relationships described above in a visual way. The top panel shows the observed characteristics, the middle panel contains the unobserved, or latent, concepts, and the bottom panel represents the outcome. Solid lines represent the measurement model, while dotted lines are intended to show the structural model. For example,

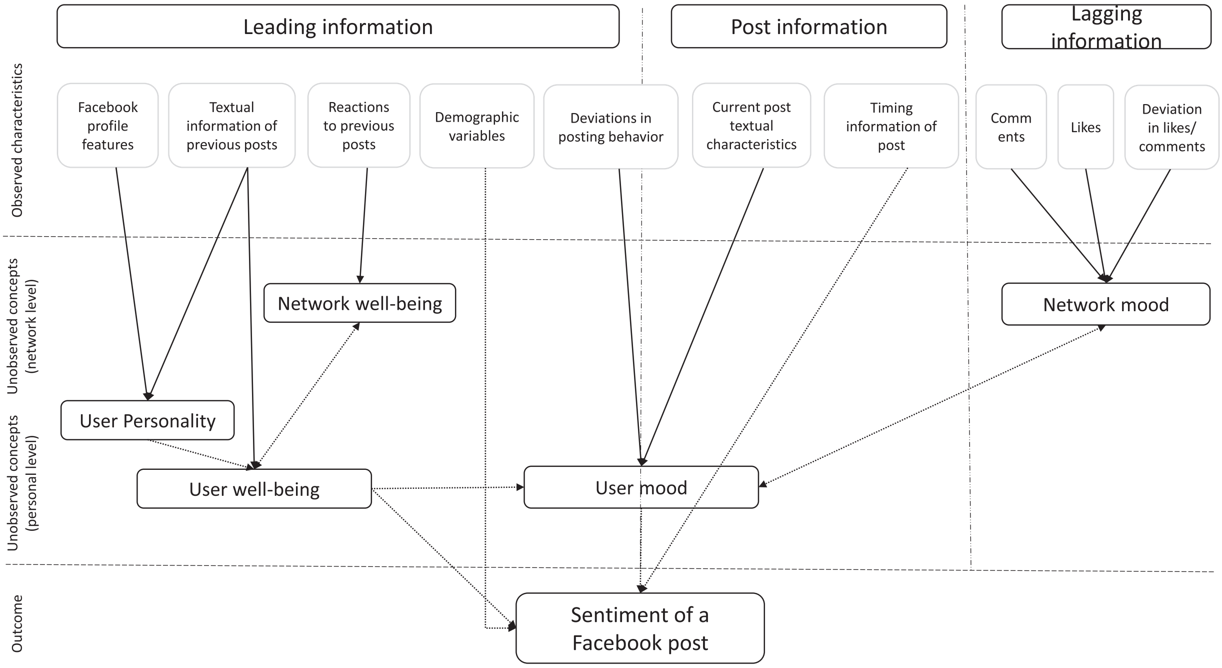

Leading information
Post information
Lagging information
Observed characteristics
Facebook profile features
Textual information of previous posts
Reactions to previous posts
Demographic variables
Deviations in posting behavior
Current post textual characteristics
Timing information of post
Comm ents
Likes
Deviation in likes/ comments
Unobserved concepts (network level)
Network well-being
Network mood
Unobserved concepts (personal level)
User Personality
User well-being
User mood
Outcome
Sentiment of a Facebook post

Fig. 1. Conceptual framework representing the literature review.

Facebook profile features are expected to, in part, measure user personality, while user personality influences user well-being and hence influences the sentiment of a Facebook post. For the sake of completeness, we also added the expected relationships for the focal post characteristics. The focal post's textual characteristics can be informative of the focal user's mood, while the timing variables are taken into account directly as control variables for post sentiment.

It is important to note that the concepts are introduced to provide plausible explanations of our findings about the relationship between the observed (top layer) characteristics and the outcome (bottom layer). Unfortunately our data do not allow us to model the concepts in the middle layer as our measurement model is incomplete. For example, there are more observed characteristics that make up the concept 'network well-being'. We do not have access to these additional characteristics and therefore it would be incorrect to make claims about that particular concept. The primary goal of our conceptual framework is to support our findings that focus on the top and bottom layer. Analysis of the middle layer is out of the scope of this research and requires additional data generated through questionnaires.

2.2. Lagging information

The lagging variables comprise information on the likes and comments of a post, as well as deviations from previous liking and commenting behavior on posts. Previous research has shown that more negatively oriented posts tend to attract more comments [87]. This can be explained by negativity bias [12]. Negativity bias is defined as the tendency to react stronger to very negative stimuli than to matched positive stimuli. In terms of engagement on posts, this means that people are more engaged with negatively oriented posts, and are willing to put more effort in commenting on the post. On the other hand, we expect that the number of likes a post receives is positively correlated with positive sentiment, as a 'like' has an inherent positive dimension. Forest et al. [37] indeed indicate that positive posts receive more likes compared to negative ones. In the case of positively oriented posts, people might simply opt to like the status, instead of taking the effort to write a comment, thereby shifting responses from comments to likes [87]. Next to the number of comments and likes, we also evaluate the valence of comments. Previous research on discussion forums and political weblogs revealed that negatively oriented posts are found to receive more negative comments, while positively oriented posts receive more positive comments [25, 49].

In accordance with the concepts of user well-being and network well-being, we propose a similar concept 'network mood', comparable to individual mood. An individual's mood can influence network mood and vice versa (e.g., by posting status updates), by mechanisms such as emotional contagion and empathy. Network mood can thus be informative about a user's mood, and hence about the sentiment of posts from that user. Since we do not have network posts available, we measure part of the network mood by the likes and comments on statuses of the focal user. As mentioned in the previous paragraph, unsuspected framings create more attention and involvement [86]. We therefore also add deviation variables, indicating if a post received more comments or likes than average for that specific user, to define network mood.

The earlier results and the theoretical framework mentioned above suggest that information on likes, comments, and deviations is very valuable to detect emotional valence of a status, and we thus hypothesize that lagging variables add predictive value to our model. The lagging variables are discussed in more detail in Section 3.4.3 and Table A3 in Appendix A. They are also shown in Fig. 1.

In sum, to the best of our knowledge we believe that this study is the first to include auxiliary features in sentiment analysis models. Based on the conceptual framework outlined in our literature review, we hypothesize that those data will significantly increase the predictive performance of our models.

3. Methodology

3.1. Data

The data were gathered using a Facebook application in the period from June 1, 2014 to July 13, 2014. The application was created for a European soccer team and advertised several times on the soccer club's Facebook page. In order to stimulate usage of our application, the users could win a jersey of the soccer team. When launching the application, the Facebook user was presented with an authorization box, which specified the data that were being collected. It was clearly stated that the data were collected solely for academic purposes. Contact information was also provided in case there were any questions. Once the user authorized the application, it started to gather personal information (e.g., gender, age, location), information on engagement behavior (e.g., Facebook groups the user belongs to, Facebook page likes, Facebook events the user attended) and general Facebook behavior (e.g., uploaded photos, videos, links and posts) from the user using the Facebook API. In total, we were able to capture 100,227 posts. As the Facebook application focused on Flemish soccer fans, the main language of the status updates is Dutch. In subsequent analyses we discard all non-Dutch posts. The average number of words used in the statuses is 15, which is comparable to the average number of words in tweets [41]. The main difference is in the maximum number of words, which goes up to 968 for our Facebook sample, while the maximum number of tweet characters is limited to 140. Detailed information about all the Facebook variables can be found in Section 3.4.2 and Appendix A.

3.2. Model description

In order to formally assess the additional value of auxiliary information over and above a focal post's content, we fit three models. The first model is the base model and reflects all the information of the present, where 'the present' refers to the time of posting (i.e., it contains the time and text variables of the post). The second model contains both information from the present and from the past by including the leading variables. The third model augments the second model with lagging variables, which adds a third time dimension to the model (i.e., the future). The choice of these three models is therefore motivated by practical reasons. We call model 1 the base model as our literature review pointed out that it reflects current practice. Model 2 has the prospect of improving predictive performance and can still be deployed in real-time. Finally, model 3 is expected to further improve performance but requires us to wait until the post has had enough time to gather comments and likes. Because model 2 can be used in real-time and model 3 cannot, it is practically relevant to determine the difference in performance between these two models. Formally, the models have the following forms:

$$\begin{aligned}
&\textit{Model 1: Status sentiment} = \textit{f(focal post's content)}\\
&\textit{Model 2: Status sentiment} = \textit{f(focal post's content)}\\
&\qquad + \textit{f(leading variables)}\\
&\textit{Model 3: Status sentiment} = \textit{f(focal post's content)}\\
&\qquad + \textit{f(leading variables)}\\
&\qquad + \textit{f(lagging variables)}
\end{aligned}$$

The definition of *Status sentiment* is described in Section 3.3, while the different independent variables are described in Sections 3.4.1, 3.4.2 and 3.4.3. The functional form of the models is not specified as we use a data mining approach without pre-set functional form, which is explained more in detail in Section 3.5.

Please cite this article as: M. Meire, et al., The added value of auxiliary data in sentiment analysis of Facebook posts, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.06.013

3.3. Dependent variable description

For the creation of our dependent variable, we follow the approach of distant supervision used by Read [80], Go et al. [41] and Pak and Paroubek [73]. This approach filters out emoticons from tweets, and uses these emoticons to represent positive and negative sentiment of a tweet. The emoticons thus serve as noisy emotion labels [41]. We list emoticons taken from Wikipedia [95] and assign a positive or negative sentiment to the emoticon. Our sentiment variable is then constructed by comparing the emoticons in the post with our reference list. In case of ties (positive as well as negative emoticons occur), the label is assigned by majority voting.

This approach implies that only Facebook messages with emoticons can be used in the training phase, which leads to a total of 17,697 available status updates (of which 2078 were classified as negative and 15,619 as positive). In order to overcome class imbalance, we apply oversampling [8].

To test the accuracy of using emoticons for sentiment detection, a random subset of 2000 status updates was manually labeled by two annotators. The inter-annotater agreement (also called Fleiss' κ [57]) between the three labels (label obtained by emoticons, annotator 1 and annotator 2) is 0.74. This score can be defined as substantial [57], which indicates that emoticons can indeed be used as sentiment labels.

3.4. Independent variable description

Different categories of variables were used in this study. As discussed above, the nature of the variables constitutes a major contribution of this paper, and hence we will further elaborate on the variables included. These can be divided into three categories: a focal post's variables, auxiliary leading variables and auxiliary lagging variables. A summary of all the variables can be found in Appendix A.

3.4.1. Focal post's variables

First, we extracted time-related variables of the post. These variables are the time, day and month of posting and a dummy variable to indicate whether the post occurred in a weekend. We include these variables as control variables [28].

Second, in order to perform the sentiment classification task, we need to process the textual information so that it can serve as input to the model. As described before, there exist a variety of text features that can be taken into account. We include as much features as possible in our predictive models, in order to have a powerful base model to test our augmented models against.

First of all, we include lexicon-based features. These features are calculated using a (Dutch) sentiment lexicon [22]. This lexicon gives a positive/negative weight to each word, as well as a subjectivity score. We then calculate the positive polarity, negative polarity, overall polarity and subjectivity for each status update by simply summing the polarity and subjectivity scores of each word in the status update. If negation words occur next to polarity words, we change the orientation of the polarity scores. These scores per status update are input features for the prediction model. Next, we use syntactic features. This includes the number of punctuations, exclamation marks, question marks, capital letters, characters and words. It also includes part-of-speech. Finally, we also create lexical features. We only include unigram features, as past research gives no conclusive evidence for the added value of higher order n-gram features [26, 41, 63, 73, 75]. In order to create the unigram, we follow the approach by Coussement and Van den Poel [23], Pak and Paroubek [73], Cao et al. [19] and D'Haen et al. [31]. In a first step, all special characters, emoticons and punctuation are removed. A tokenization is performed by splitting each status in distinct words using spaces as separators. Next, stopwords such as 'the' or 'a' are removed since these words are frequently used and hold little or no content information [38]. Abbreviations are replaced using a dictionary and a spelling check is conducted in order to cope with the noisy nature of social media data. Indeed, users often use their cell phones to post status updates which leads to a higher frequency of misspellings and slang [41, 72]. The next step is lemmatization, followed by synonym replacement in order to further reduce the vector space. As a final step, stemming is applied. With stemming, a word is stripped to the basic form (i.e., suffixes and prefixes are removed) [31, 54, 77]. This process results in a basic unigram (also called bag-of-words or document-term matrix). The unigrams obtained by the procedure described above are still very sparse. Therefore, we apply a feature selection technique that reduces the number of features for input to the classification algorithm. We chose to work with Latent Semantic Indexing (LSI). This method is proposed by Deerwester et al. [29] and reduces the original matrix in dimension by its first k principal component directions [29].

3.4.2. Leading variables

Leading variables can be subdivided into five groups, as outlined in Section 2. Facebook profile features contain engagement behavior (e.g., number of Facebook events attended) and general Facebook behavior (e.g., number of photos, videos). Age and gender are included as demographic variables. Previous post information will control for the user's and network well-being. This information includes average measures, e.g. average polarity of posts and average number of likes on previous posts. Deviations from previous post information can be informative about user mood. We use the following equation to calculate these deviation variables:

$$\Delta X_{i,T} = X_{i,T} - \bar{X}_{i,1\rightarrow T} \tag{1}$$

where X denotes the specific variables, i represents a user and T indicates the time of posting. We thus calculate for every post the deviation between the post's feature score and the average feature score for the user that posted. Example variables are the deviation in the number of words and the deviation in the number of positive and negative words of the post. A complete list can be found in Table A2 in Appendix A.

3.4.3. Lagging variables

Lagging variables can only be observed after the content was posted, which are, in the case of Facebook, likes and comments. We thus include the number of likes, the number of comments, the number of likes on comments and textual information from comments (e.g., the number of positive or negative words in comments, the number of words in comments) into our predictive model. Furthermore, as for the leading variables, we calculate deviations from the normal liking or commenting behavior on posts of the focal user. This includes for example the deviation in the number of comments and the deviation in the number of likes. In order to calculate the lagging variables, we allow each post to gather likes and comments for 7 days. We chose this particular time frame for three reasons. First, this limitation increases the practical feasibility of our solutions as sentiment analysis is most valuable within a short time frame. Second, as such we give each post equal time to gather likes and comments. Third, as Fig. 2 shows, more than 99% of all comments are gathered during the first week. A complete list of all lagging variables can be found in Table A3 in Appendix A.

3.5. Predictive techniques

We use the Support Vector Machines (SVM) and Random Forest classification algorithms to perform our sentiment analysis. SVM has been used extensively in sentiment analysis and generally outperforms other methods such as Naïve Bayes, Maximum Entropy and logistic regression [35]. Although Random Forest classification has

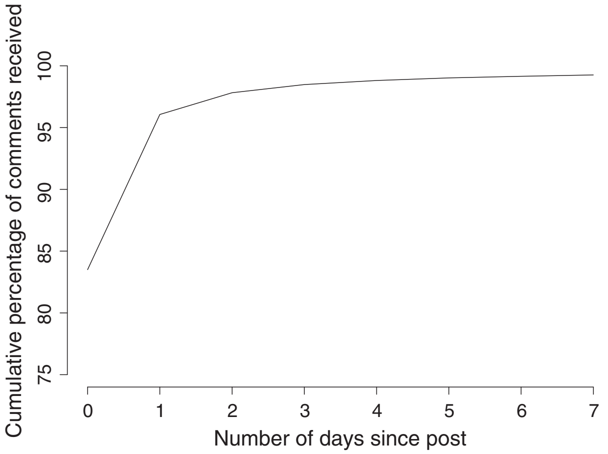


Fig. 2. Cumulative collected % of comments per day.

not been frequently used in sentiment analysis, it has recently been shown to be the best allround classification technique in many other domains [35]. Using both algorithms allows to use a well-established technique in sentiment analysis on the one hand, while on the other hand we can assess whether the Random Forest classification algorithm adds value in sentiment analysis.

3.5.1. *Support Vector Machines*

An important parameter in SVM is the kernel function [11]. We use a radial basis (RBF) kernel, because this allows for non-linear relationships and requires the choice of only one hyperparameter γ, the width of the Gaussian [14]. We thus have, combined with the SVM penalty parameter C, two parameters to choose. The choice of these parameters cannot be determined in advance. Hence, we follow the recommendation to test different values of C, ($C = [2^{-5}, 2^{-4}, \ldots, 2^{15}]$) and γ, ($\gamma = [2^{-15}, 2^{-14}, \ldots, 2^{3}]$) [48]. We use the svm function of the *e1071 R* package [66] to implement SVM.

3.5.2. *Random Forest*

The Random Forest classification algorithm grows a committee of classification trees and averages over all tree predictions [18]. By doing so, it can overcome the limited robustness and suboptimal performance of individual trees [34]. Applying Random Forest has multiple advantages. It is does not overfit [18]. Furthermore, it is easy to use in that variable importances are provided [82] and only two parameters have to be set [16]: the number of trees and the number of predictors to consider at each step in the tree. We set these parameters according to the guidelines of Breiman [18]: the number of trees is set to 1000 and the number of predictors is defined as the square root of the total number of variables. Random Forest is implemented using the *randomForest* package in *R* provided by Liaw and Wiener [59].

3.6. *Performance evaluation*

Instead of classifying each post with a binary label {negative, positive}, we compute a score, representing the probability that a post is positive. For example, instead of saying that a post is positive, we would be able to say that the post is 70% likely to be positive, which is equivalent to saying that the post is 70% positive. Therefore, model performance is measured by the area under the receiver operating characteristic curve (AUC or AUROC). In case of scoring classifiers the AUC is a more adequate performance measure than, for example, accuracy as it does not rely on the cut-off values of the posterior probabilities [6]. AUC is defined as follows:

$$AUC = \int_0^1 \frac{TP}{(TP+FN)} d \frac{FP}{(FP+TN)} = \int_0^1 \frac{TP}{P} d \frac{FP}{N} \tag{2}$$

with TP: True Positives, FN: False Negatives, FP: False Positives, TN: True Negatives, P: Positives (positive sentiment), N: Negatives (negative sentiment). The values of the AUC range from 0.5 to 1. An AUC of 0.5 means that the model is not able to do better than a random selection, while a value of 1 indicates a perfect prediction [6].

3.7. *Cross validation*

We use five times twofold cross-validation (5x2 CV) [3, 33]. This method randomly splits the sample into two partitions of equal size. The first partition serves as training set while using the second partition as test set and vice versa. This procedure is repeated 5 times. Hence, a total of 10 performance measures per model will be obtained [33]. We summarize these 10 performance measures with the median. To assess whether the AUCs of the different models are significantly different, we use the non-parametric Friedman test [39] as suggested by Demšar [30]. The models are ranked, per fold separately, with the best model receiving the rank of 1, the second receiving the rank of 2 and the worst performing model receiving the rank of 3. In case of ties, the average rank is assigned. The Friedman statistic can then be defined as:

$$\chi_F^2 = \frac{12N}{k(k+1)} \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right] \tag{3}$$

where N is the number of folds, k is the number of models and R_j is the average rank of the j-th model over all folds.

3.8. *Variable importance measures and Partial Dependence Plots*

In order to interpret the relationships between independent variables and the sentiment classification, we will use the Random Forest models. The variable importances are assessed using the total decrease in node impurities from splitting on the variable, averaged over all trees in the Forest. The node impurity is measured by the Gini index $p(1-p)$, and the decrease in node impurity is measured as follows:

$$\Delta(s,\tau) = p_\tau(1-p_\tau) - \left(\frac{|\tau_L|}{|\tau|} p_{\tau_L}(1-p_{\tau_L}) + \frac{|\tau_R|}{|\tau|} p_{\tau_R}(1-p_{\tau_R}) \right) \tag{4}$$

where s is short for a given split of a given variable and τ, τ_L, τ_R respectively stand for all the cases in the parent node, left child node and right child node. p is short for $p(y=1)$ with $y = \{0,1\}$ and thus denotes the probability that an observation is positive given that it is in that specific node. We denote cardinality by $|\bullet|$. We use the importance function in the *randomForest* package in *R* [59]. Remark that we take the median of the five times twofold cross-validated mean decrease in node impurity when we report importance measures.

Next to the most important variables, we are interested in the form of the relationship between predictors and the response. For this purpose, we use Partial Dependence Plots [44]. Partial Dependence Plots can be used to interpret any 'black box' model. Basically, the plots represent the relationship of one (or a subset) of the predictors with the response, taking into account the effect of all the other predictor variables. The Partial Dependence Plots are five times twofold cross-validated, using the *interpretR R* package [7].

Please cite this article as: M. Meire, et al., The added value of auxiliary data in sentiment analysis of Facebook posts, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.06.013

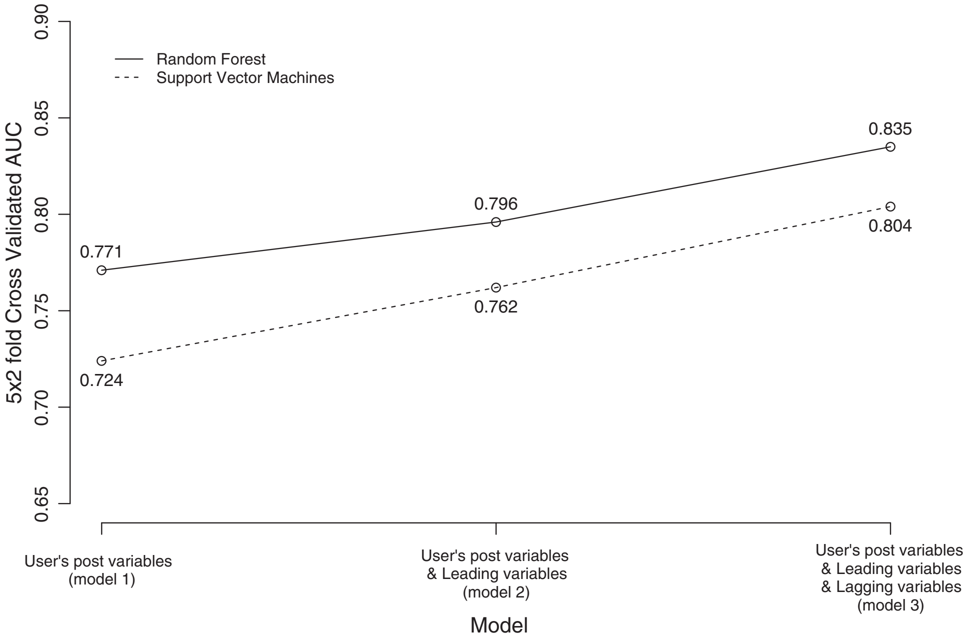


Fig. 3. Results of the models in terms of AUC.

4. Discussion of results

As explained in Section 3.2, three models were built. The first model only considers the present information, the second model considers both present and past information and the third model considers present, past and future information. Fig. 3 shows the performance of the three models in terms of AUC, both for the Random Forest (solid line) and Support Vector Machine (dashed line) models. As the Random Forest algorithm creates better models across the board, all subsequent results will be discussed in terms of the Random Forest model. Remark that the reported AUCs are median values of the five times twofold cross-validation procedure.

The Friedman test indicates the presence of a significant difference in the analysis ($\chi^2_3 = 20, p < 0.01$). Subsequently we made pairwise comparisons between the models and found that on each of the ten folds, the second model performs better that the first model, and the third model performs better than the second model. This means that model 2 is significantly better than model 1 (p=0) and that model 3 is significantly better than model 2 (p=0).

In sum, the AUCs show that leading and lagging variables add value to the user's post variables. In order to understand what drives these results, we analyzed the variable importances. The top 50 variable importances of the best, most comprehensive model (the third Random Forest model: user's post variables & leading variables & lagging variables) are shown in Fig. 4 and listed in Appendix B. In Fig. 4, the variables are sorted in descending order of (5x2 CV median) mean decrease in Gini, which means that the most important variables are ranked first. When looking at the graph, we see that the top 10 importances are mixed among the three components of the model; three variables originate from the user's post variables, three

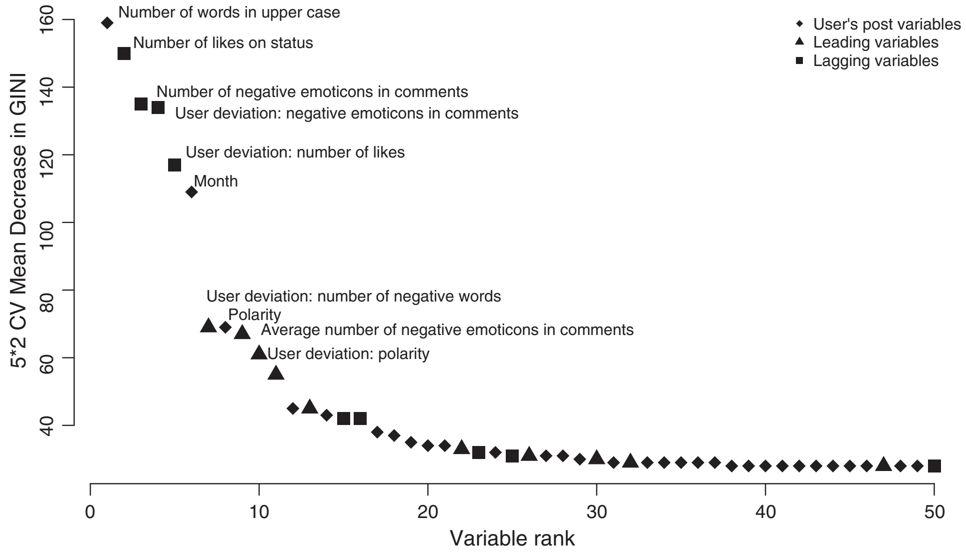
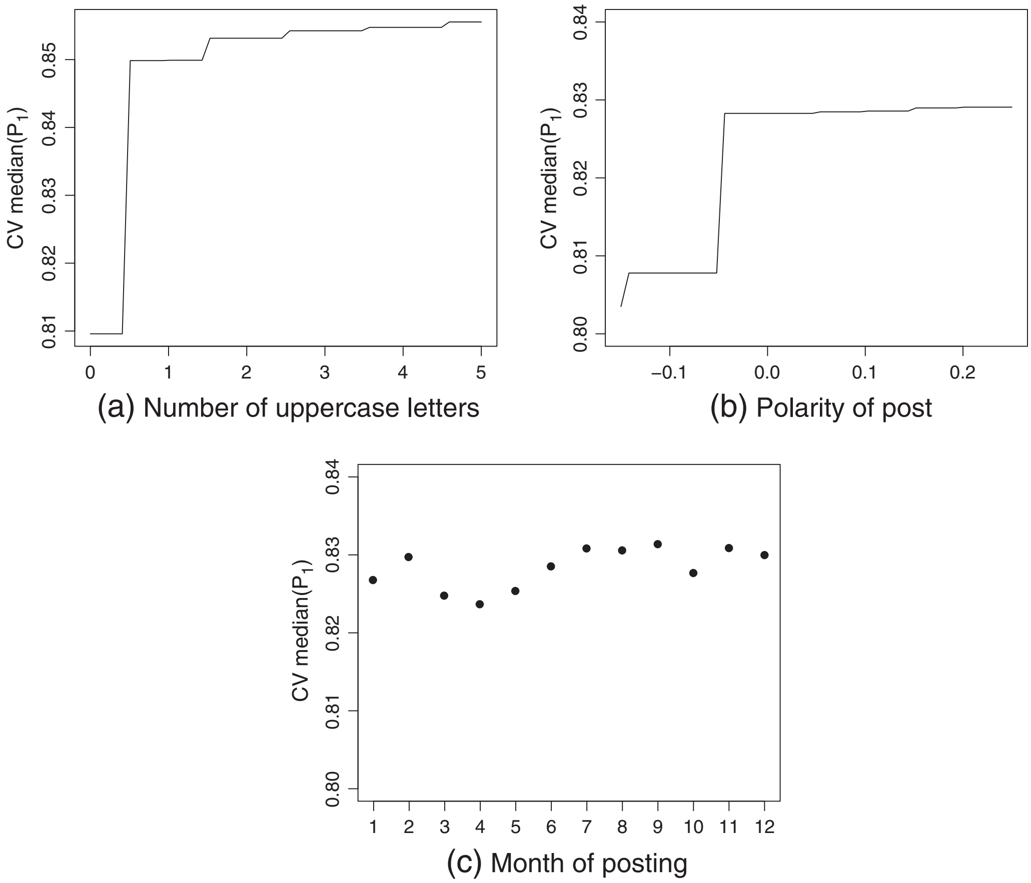


Fig. 4. Variable importances of most complete model.



(a) Number of uppercase letters

(b) Polarity of post

(c) Month of posting

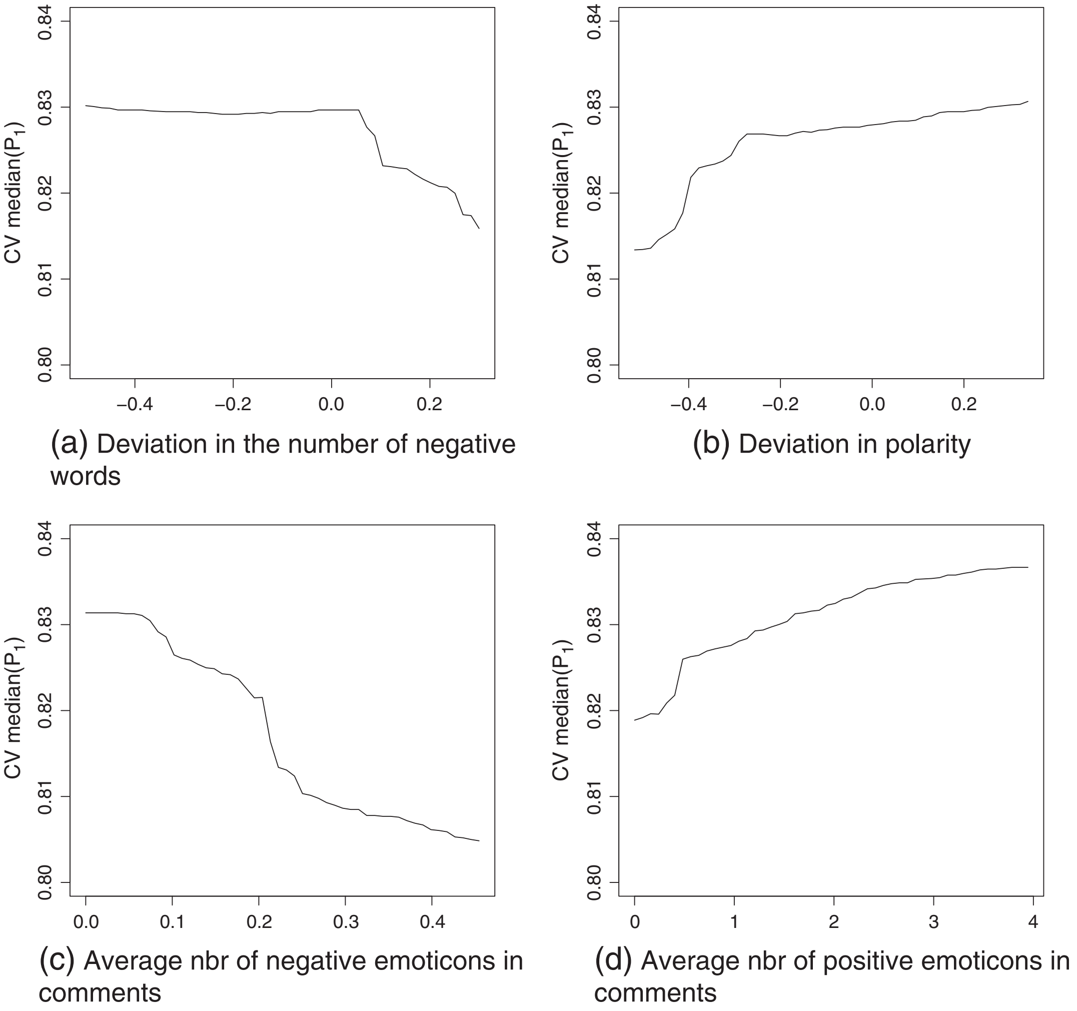
Fig. 5. Partial Dependence Plots of post variables.

variables are leading variables and the remaining four variables contain lagging information. This again suggests that all data sources are complementary to each other. We will continue with a discussion of the top post, leading, and lagging variables, starting with the post variables. We use Partial Dependence Plots (PDPs) for this purpose. The PDPs depict the predicted probability of a positive post on the y-axis, and the different values of the predictor on the x-axis.

We see that the number of uppercase letters and the post polarity both have a positive relationship with positive sentiment, as depicted in Figs. 5a and 5b. For polarity, this was expected as it measures the positivity of a post based on the lexicon approach. Our research also suggests that the number of uppercase letters is strongly related to positive sentiment. Capital letters are used when users are more passionate about the post. They are often used as intensifiers of the message [88]. A look at the negative posts in our sample brings up a possible explanation for the positive direction of the intensifier. Negative posts on Facebook frequently convey low-arousal negative feelings (e.g., feeling sick, alone) instead of high-arousal feelings such as complaints or anger. This means that there is no need to use intensifiers for these negative feelings, leaving intensifiers to be used mainly for positive posts. Although several papers include uppercase words or letters as features, none of the papers report the importance of the uppercase feature separately, making it impossible to compare our results. Finally, month of posting is an important predictor. The plot (Fig. 5c) does not show a clear pattern, except that spring months score a little bit lower than average. This can be caused by the relatively poor performance of the soccer team during this period. Indeed, a larger proportion of the posts is related to this soccer team compared to a completely random selection of posts. As such, this result is not immediately generalizable, but we show the importance of including timing variables as control variables in sentiment analysis. Finally, it is worth noting that Appendix B shows that 30 out of the top 50 variables are post variables.

Fig. 6 shows the Partial Dependence Plots for the leading variables. The deviation in the number of negative words and polarity are shown in the top row (Fig. 6a and 6b). A higher deviation in the number of negative words (i.e., more negative words are used than on average) leads to a higher probability of negative sentiment. A negative deviation in polarity leads to a higher probability of negative sentiment as well. This means that if the polarity of a post is more negative than the user's average post, the post will receive a more negative score. Fig. 6c and 6d shows the average number of negative/positive emoticons in comments (the average number of positive emoticons in comments is the eleventh most important variable). We see that a higher average number of positive/negative emoticons in comments on previous posts, indicates a higher probability of a positive/negative focal post. This supports our conceptual framework and indicates that well-being can be predictive of sentiment. Furthermore, Fig. 6a and 6b indicate that also mood, as a temporal change of subjective well-being, can be informative. Indeed, Ortigosa et al. [71] state that behavior variations, such as deviations from the average polarity of posts shown in Fig. 6a and 6b, indicate changes in the user's mood. Finally, when looking a the top 50 most important variables, we see age as an important demographic variable, and the mean and standard deviation of the time between the focal user's page likes as important personality-related variables.

Please cite this article as: M. Meire, et al., The added value of auxiliary data in sentiment analysis of Facebook posts, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.06.013



(a) Deviation in the number of negative words

(b) Deviation in polarity

(c) Average nbr of negative emoticons in comments

(d) Average nbr of positive emoticons in comments

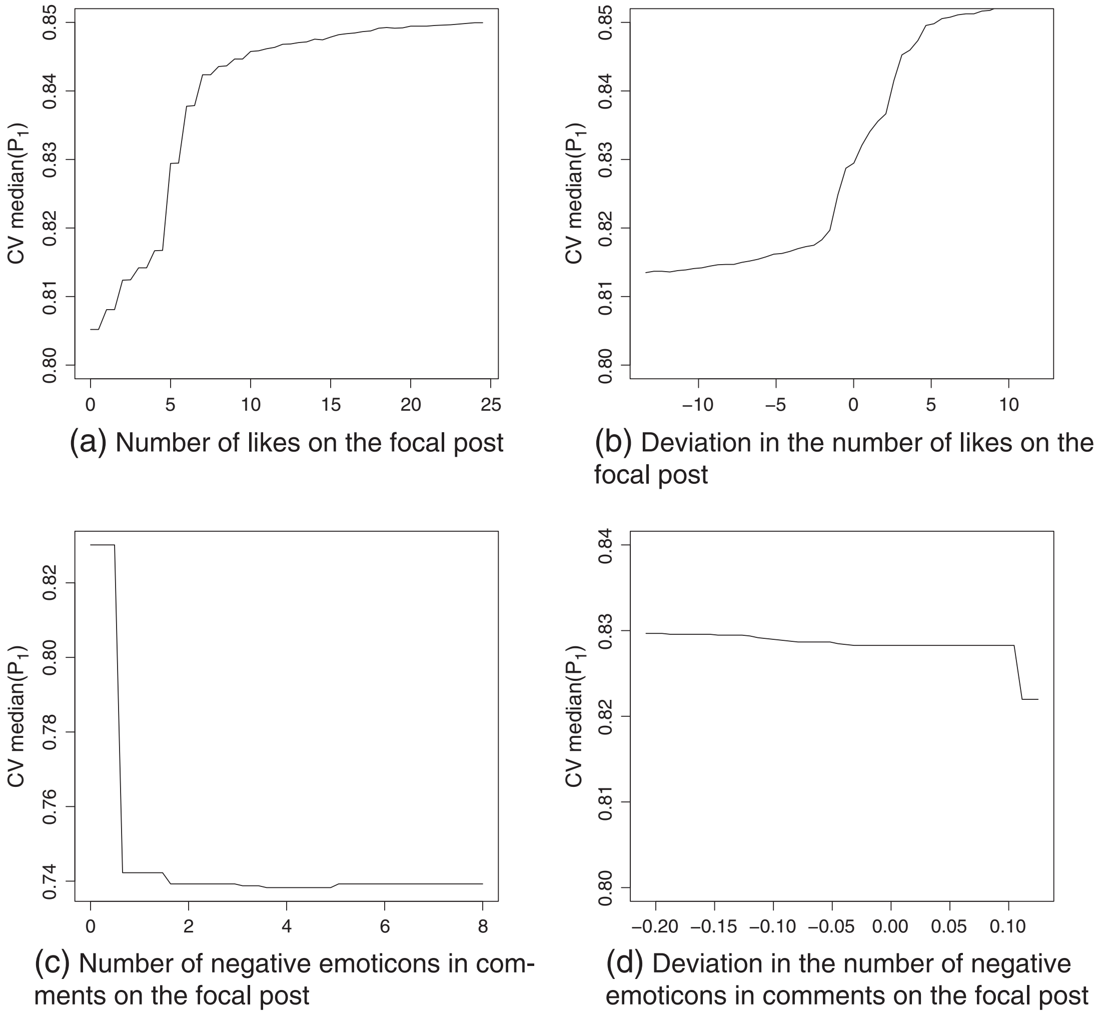
Fig. 6. Partial Dependence Plots of main leading variables.

Finally, the top lagging variables are discussed. These are plotted in Fig. 6. While the number of likes (depicted in Fig. 7a) are very predictive, the number of comments do not seem that important (only fiftieth most important variable; not shown). The relationship of likes is as expected: the higher the number of likes, the higher the probability of positive sentiment. Fig. 7b shows the deviation in the number of likes compared to the average number of likes on posts by the same user. If the post receives less likes compared to an average post, the probability of positive sentiment declines. Fig. 7c and 7d show the number and deviation of negative emoticons in comments on the focal post. Both graphs shows that a higher number of negative emoticons, both in absolute figures and compared to the average number of the user, indicate a higher probability of negative sentiment. These results confirm the earlier findings of Stieglitz and Dang-Xuan [87], and also support our conceptual framework relating to network mood, user mood and post sentiment. Stieglitz and Dang-Xuan [87] also found a positive relationship between positive emoticons in comments and the positive sentiment of a post. We find this variable on a sixteenth place, with indeed a positive relationship (not shown), but of much smaller magnitude.

All previous results apply to a model trained and tested on posts with emoticons, which are used as noisy labels. These posts may be easier to predict than regular posts, because they express clear and strong emotions. Therefore, we manually labeled a random sample of 2000 posts without emoticons, and tested the model on these posts. The inter-annotator agreement (Fleiss' κ) for the statuses is 0.81, indicating that the task was well-defined [57]. The annotators disagreed in 198 cases, which were subsequently revised and assigned a final sentiment label in order to include them in the analysis. For subsequent analysis, we dropped neutral statuses (259 cases) [26, 41, 75]. In that way, we can apply our model to the new statuses, which are used as new test samples for each of the folds. Results showed that model 1 achieved a median AUC of 0.751, model 2 a median AUC of 0.775 and model 3 a median AUC of 0.812. We can conclude that (1) the focal post's variables show significantly lower performance compared to models using statuses with emoticons, probably because emotions are expressed less clearly and (2) there is an effect of both leading and lagging variables. The effects in terms of extra predictive power are very similar to the case of statuses with emoticons. In summary, the results for posts with and without emoticons are very similar and consistent in terms of the added value of leading and lagging information.

5. Conclusion and practical recommendations

Initially, sentiment analysis was performed mainly on review data. Recently, because of their abundance, social media data have become the main focus in the field. Despite this change in focus, our literature review shows that researchers have not yet explored the additional wealth of information that is available through social media data. Therefore, in this study we set out to (1) study the added



(a) Number of likes on the focal post

(b) Deviation in the number of likes on the focal post

(c) Number of negative emoticons in comments on the focal post

(d) Deviation in the number of negative emoticons in comments on the focal post

Fig. 7. Partial Dependence Plots of main lagging variables.

value of leading and lagging variables for sentiment analysis, (2) determine the top predictors, (3) and explore the relationships of the top predictors with the sentiment of a post. We devised a conceptual framework to support our results.

The results clearly indicate that leading and lagging variables add predictive value to established sentiment analysis models. In other words, past and future information does add value over present information. The magnitude of the differences in model performance and the consistency of these differences over all folds suggest that the results are relevant. Given that Facebook messages are informal and therefore often contain slang, irony or multi-lingual words [72], sentiment analysis is difficult based solely on text. We showed that leading and lagging variables can help to predict sentiment in this challenging environment, and our conceptual framework helped in explaining why these variables matter.

The most important predictors of the most complete model were a mix of post variables (e.g., number of uppercase letters), leading variables (e.g., average number of negative comments on posts in the past) and lagging variables (e.g., number of likes) indicating that all three model components add to the predictive value of our model. We can draw several conclusions from these findings.

First, we can see that word use and time of posting are important. The number of uppercase letters is the most important predictor, followed by month of posting and the use of negative and positive words (polarity) as the sixth and eighth most important factors, respectively. Moreover, we see that a deviation in polarity is important, indicating a mood change from the general subjective well-being of the user, thereby supporting our predictions based on the conceptual framework. Finally, in total 30 of the 50 most important variables are related directly to the post's content and time of posting.

Second, it becomes clear that reactions on status updates contain relevant information, as 6 out of the 10 most important predictors stem from likes and comments related variables. A higher number of likes indicates a more positive post, while negative emoticons in the comments (on the current post, on previous posts, and deviations from previous posts) indicate negative posts. It thus seems that there is additional information in the variables that measure network well-being and mood. This also confirms previous findings from Stieglitz and Dang-Xuan [87].

Third, we can conclude that general Facebook variables and demographics seem less important. Age is the thirteenth most important variable, while only two Facebook-related variables show up in the top 50 (the average and standard deviation in page liking behavior of the user). Page liking behavior has already been shown to be predictive of, among others, happiness and personality traits [52], and thus user well-being, which makes this result plausible. The implication is that one could save the burden to gather the immense amount of data from Facebook, as the majority of the variables have only limited importance. Based on our results, we thus argue that

Please cite this article as: M. Meire, et al., The added value of auxiliary data in sentiment analysis of Facebook posts, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.06.013

age, page liking behavior and of course posts of the user are the most important Facebook variables to identify.

Finally, we would like to make a general remark on the importance of variables. We see that negative variables receive more attention from the algorithm than positive variables, or that deviations in the negative direction have a bigger influence. This can be linked to the lower number of negative posts in our sample and on Facebook in general [60, 70]. As the majority of the posts is positive, clues about negative sentiment turn out to be, in general, more useful to the algorithm. Therefore, we conclude that in a setting where the ratio of positive versus negative posts is high, features that indicate negativity can be more helpful to predict overall sentiment.

Academics, companies and public parties are interested in large scale sentiment analysis, which yields a wide range of applications. Companies can perform sentiment analysis to analyze customer satisfaction [41], to increase ad-targeting efforts or to track public opinion about the company. Teachers can use sentiment analysis to support personalized e-learning [72]. Academics measure general public mood and track changes over time. Political parties employ social media to track public sentiment and adjust their campaign towards regions or topics that suffer from negative emotions. Finally, broadcasters and media can analyze tweets to predict election outcomes [92].

Established approaches to sentiment analysis described above include only present information. We propose to include all information from the past, which includes previous posts from the same user, in any sentiment analysis model. Indeed, even real-time applications can include leading information and benefit from the extra predictive value. Live television, for example, can analyze reactions on the Facebook or Twitter page in real-time, thereby including leading information. Another example may be news channels that analyze tweets real-time to predict elections (e.g., on the day of election), thereby using leading information. This could enable a more accurate prediction and better reputation of the news channel. On the other hand, real-time applications cannot benefit from lagging variables. However, other applications can take advantage of these lagging variables. For example, a company can allow for a small lag in the measurement of customer satisfaction. This study used a lag of 7 days, but as Fig. 2 shows, more than 95% of all comments are gathered after only one day. The time frame for creating the lagging variables can thus be shortened, without losing much of the information. Finally, one can use the present and past information in a first round to quickly get an idea of the sentiment, and refine these early findings with lagging information in a second round. One possible application is a marketing campaign for a new product. First, the company can perform sentiment analysis to assess global sentiment concerning the product. In this way, the broad outlines of the marketing campaign can be adjusted if necessary. Second, more fine-grained sentiment analysis, including lagging variables, can be performed that allows to fine-tune the campaign. In sum, we feel that our proposed approach is a promising path for many sentiment analysis applications.

6. Limitations and future research

Sentiment analysis can be applied to a wide range of sources. Our research shows that leading and lagging information can be very valuable in the context of sentiment analysis on Facebook posts. It remains unclear whether a similar approach can work for other media such as Twitter and review data, but we argue that the central idea is generalizable. Indeed, Twitter also includes leading information such as a concise user profile and previous tweets, while retweets and favorites can be seen as lagging information embedded in Twitter. An interesting avenue for further research would thus be to extend the application to other social media platforms.

Although our study extends the use of data that is available in social media to predict sentiment, and includes emotional contagion to some extent, we did not include complete network information in the analysis. Network effects are, to the best of our knowledge, not yet discussed in the area of sentiment analysis. However, there is a growing amount of research on social networks reporting the importance of network effects on a wide range of behaviors (e.g., Bakshy et al. [5]). As the main drivers of these effects are homophily and social influence [43], it can be expected that a user's emotions are related to the emotions of a user's network. Further research could try to incorporate network data and improve our results.

The third direction for future research is to use a more theoretical angle to approach the problem, while our primary goal was to look at the added value of leading and lagging variables taking a data mining approach. With the current results, it can be interesting to take a look at the underlying constructs of (individual and network) well-being, mood and personality, and incorporate these constructs rather than all Facebook variables separately (e.g., by using a questionnaire). In this study we use latent constructs to provide plausible explanations of our findings about the relationship between the observed characteristics and the outcome variable, sentiment. As mentioned in the literature review, our data do not allow us to model the latent constructs as our measurement model is incomplete. We work with observed data and retrofitted latent constructs on these variables. Future research could start from latent constructs and make sure appropriate variables are included to fully measure each construct, which would allow for a formal measurement model. A logical approach would be to use data generated through surveys and use appropriate measurement scales. Because this study uses observed data we are unable to sort this out. Nevertheless, the unobserved concepts allow us to strengthen the theoretical underpinnings of our study, and facilitate the discussion of our results. We also feel that our conceptual model is a good basis for future theoretical and empirical research.

The fourth limitation is selection effects. It might be possible that the users whose information was obtained by using the application may be different from users that did not use the application. The Facebook application was developed for a European soccer team, which means that the users of the application are interested in soccer. This can also have its repercussions on the posts that are analyzed (i.e., they may be more soccer-oriented than the average Facebook post). In our opinion, this does not impose serious repercussions on the obtained results. In the case the posts are more biased towards one domain (e.g., soccer), it is likely that the text variables become more predictive because posts are more related and that sentiment is easier to predict [72]. In this context, we were able to substantially improve our predictions by adding leading and lagging information. In case the domain is less bounded, it is likely that leading and lagging information can have even more predictive value.

The fifth limitation of this study is the limited number of values that some of the variables can have. Facebook limits the number of occurrences of a variable (e.g., the likes of a user) to the 25 most recent entries. This issue is most important for frequency variables that are included as part of the user profile information (which is part of the leading information). In order to deal with this limitation, we calculated frequency within a specific period of time. The length of this time window per variable is determined as to no user in our database reaches the maximum number of 25 entries.

As a final remark we want to say that although this study has some shortcomings, it is the first sentiment analysis study using such a variety of data. We feel that this is a valuable contribution to literature.

Appendix A. Variable list

Table A1
Focal post's variables.

Variable name	Variable description (Category)
SVD concept 1 - 100	SVD concepts (Lexical)
Number_uppercase	Number of uppercase letters in post (Lexical)
Number_punct	Number of punctuations in post (Lexical)
Number_qm	Number of question marks in post (Lexical)
Number_em	Number of exclamation marks in post (Lexical)
Number_nbr	Number of numbers in post (Lexical)
Number_wow	Number of ‘wow’ (or similar like ‘woooow’) mentioned in post (Lexical)
Number_pf	Number of ‘Pf’ (or similar like ‘Pffff’) mentioned in post (Lexical)
Number_lol	Number of ‘lol’ mentioned in post (Lexical)
Number_characters	Number of characters in post (Lexical)
Number_words	Number of words in post (Lexical)
Number_pos_words	Number of positive words in post (Lexicon)
Number_neg_words	Number of negative words in post (Lexicon)
Positive_polarity	Sum of positive polarity scores for the post (Lexicon)
Negative_polarity	Sum of negative polarity scores for the post (Lexicon)
Polarity	Sum of polarity scores for the post (Lexicon)
Subjectivity	Sum of subjectivity scores for the post (Lexicon)
POS_noun	Number of nouns in post (Syntactic)
POS_verb	Number of verbs in post (Syntactic)
POS_adj	Number of adjectives in post (Syntactic)
Month	Month of post (Time)
Weekday	Day of week of post (1 to 7) (Time)
Weekend	Dummy indicating if post occurred during weekend (Time)
Time_of_day	Time of the day of post (Time)

Table A2
Leading variables.

Variable name	Variable description (category)
Previous post information	
Mean_neg_emo	Average number of negative comments received on previous posts
Mean_pos_emo	Average number of positive comments received on previous posts
Mean_likes_posts	Average number of likes received on previous posts
Mean_comm_posts	Average number of comments received on previous posts
Mean_comm_likes_user	Average number of comments received on previous posts, liked by the user
Total_nbr_likes	Total number of likes received on previous posts
Total_nbr_comments	Total number of comments received on previous posts
Mean_polarity	Mean polarity of previous posts
Mean_pos_words	Mean number of positive words in previous posts
Mean_neg_words	Mean number of negative words in previous posts
Mean_subjectivity	Mean subjectivity of previous posts
Mean_nbr_words	Mean number of words in previous posts
Deviation_polarity	Deviation in polarity of the focal status compared to previous posts
Deviation_pos_words	Deviation in number of positive words in the focal status compared to previous posts
Deviation_neg_words	Deviation in number of negative words in the focal status compared to previous posts
Deviation_subjectivity	Deviation in subjectivity of the focal status compared to previous posts
Deviation_nbr_words	Deviation in number of words in the focal status compared to previous posts
Total_nbr_posts	Total number of previous posts
General Facebook information	
Age	Age of user (personal information)
Gender	Gender of user (personal information)
Relationship_single	Dummy indicating whether the person is in a relationship or not (personal information)
General Facebook information	
Heterosexual	Dummy indicating whether the person is heterosexual (personal information)
Account_age	Age of the Facebook account of the user (personal information)
Number_friends	Number of friends of the user (personal information)
Number_groups	Number of Facebook groups the user is member of (engagement behavior)
Number_likes	Number of Facebook pages the user has liked (engagement behavior)
Number_events	Number of Facebook events the user had indicated to attend (engagement behavior)
Number_interests	Number of interests as expressed on Facebook by the user (engagement behavior)
Number_check-ins	Number of check-ins registered on Facebook (engagement behavior)
Number_cin_likes	Number of likes on check-ins (engagement behavior)
Number_cin_tags	Number of tags related to check-ins (engagement behavior)
Number_cin_comments	Number of comments related to check-ins (engagement behavior)
Number_photos	Number of photos (general FB behavior)
Number_videos	Number of videos (general FB behavior)
Number_links	Number of links (general FB behavior)
Number_posts	Number of posts (general FB behavior)
Number_comm_photos	Number of comments received on photos (general FB behavior)
Number_comm_videos	Number of comments received on videos (general FB behavior)
Number_comm_links	Number of comments received on links (general FB behavior)
Number_likes_photos	Number of likes received on photos (general FB behavior)
Number_likes_videos	Number of likes received on videos (general FB behavior)
Number_likes_links	Number of likes received on links (general FB behavior)
Recency_comment	Recency of comments received from other users (general FB behavior)
Recency_likes	Recency of likes received from other users (general FB behavior)
Recency_photo	Recency of last photo at time of post posting (general FB behavior)
Recency_video	Recency of last video at time of post posting (general FB behavior)
Recency_link	Recency of last link at time of post posting (general FB behavior)
Recency_check-in	Recency of last check-in at time of post posting (general FB behavior)
Recency_like	Recency of last page like at time of post posting (general FB behavior)
Recency_post	Recency of last post at time of focal post (general FB behavior)
Mean_time_photos	Average time between photo uploads (general FB behavior)
Mean_time_videos	Average time between video uploads (general FB behavior)
Mean_time_links	Average time between links (general FB behavior)
Mean_time_likes	Average time between user likes (on pages) (general FB behavior)
Mean_time_posts	Average time between post (general FB behavior)
SD_time_photos	Standard deviation of the time between photo uploads (general FB behavior)
SD_time_videos	Standard deviation of the time between video uploads (general FB behavior)
SD_time_links	Standard deviation of the time between links (general FB behavior)
SD_time_likes	Standard deviation of the time between user likes (on pages) (general FB behavior)
SD_time_posts	Standard deviation of the time between post (general FB behavior)
Profile_completeness	Number of Facebook profile items filled in by the user (general FB behavior)

Table A3
Lagging variables.

Variable name	Variable description
Nbr_likes	Number of likes the focal post received in 7 days
Nbr_comments	Number of comments the focal post received in 7 days
Nbr_own_comm	Number of comments made on the focal post by the focal user
Nbr_comm_persons	Number of persons commenting on the focal post
Nbr_comm_likes	Number of likes on comments received on the focal post
Nbr_words_comm	Number of words in the comments received on the focal post
Nbr_punct_comm	Number of punctuations in comments received on the focal post
Nbr_qm_comm	Number of question marks in comments received on the focal post
Nbr_em_comm	Number of exclamation marks in comments received on the focal post
Nbr_upper_comm	Number of uppercase letters in comments received on the focal post
Nbr_lol_comm	Number of ‘lol’ mentioned in comments received on the focal post
Neg_emo_comm	Number of negative emoticons in comments received on the focal post
Pos_emo_comm	Number of positive emoticons in comments received on the focal post
Dev_nbr_likes	Deviation in the number of likes received on the focal post compared to previous posts
Dev_nbr_comments	Deviation in the number of comments received on the focal post compared to previous posts
Dev_nbr_own_comm	Deviation in the number of own comments made on the focal post compared to previous posts
Dev_nbr_comm_persons	Deviation in the number of commenting persons on the focal post compared to previous posts
Dev_nbr_comm_likes	Deviation in the number of likes received on comments on the focal post compared to previous posts
Dev_neg_emo	Deviation in the number of negative emoticons in comments received on the focal post compared to previous posts
Dev_pos_emo	Deviation in the number of positive emoticons in comments received on the focal post compared to previous posts
Comments_span	The time span in which comments were received

Appendix B. Variable importance scores

Table B1
Variable importances (top 50).

Rank	5*2 CV median mean mecrease in Gini	Variable name	Category
1	159	Number_uppercase	Focal post’s variables
2	150	Nbr_likes	Lagging variables
3	135	Neg_emo_comm	Lagging variables
4	134	Dev_neg_emo	Lagging variables
5	117	Dev_nbr_likes	Lagging variables
6	109	Month	Focal post’s variables
7	69	Deviation_neg_words	Leading variables
8	69	Polarity	Focal post’s variables
9	67	Mean_neg_emo	Leading variables
10	61	Deviation_polarity	Leading variables
11	56	Mean_pos_emo	Leading variables
12	45	Number_punctuation	Focal post’s variables
13	45	Age	Leading variables
14	43	Number_neg_words	Focal post’s variables
15	42	Dev_nbr_comments	Lagging variables
16	42	Dev_pos_emo	Lagging variables
17	38	SVD Concept 1	Focal post’s variables
18	37	SVD Concept 22	Focal post’s variables
19	35	Weekday	Focal post’s variables
20	35	SVD Concept 29	Focal post’s variables
21	34	SVD Concept 2	Focal post’s variables
22	33	Mean_likes_posts	Leading variables
23	32	Nbr_comm_persons	Lagging variables
24	32	SVD Concept 62	Focal post’s variables
25	31	Nbr_words_comm	Lagging variables
26	31	Total_nbr_likes	Leading variables
27	31	SVD Concept 21	Focal post’s variables
28	31	SVD Concept 28	Focal post’s variables
29	30	SVD Concept 99	Focal post’s variables
30	30	Mean_time_likes	Leading variables
31	29	SVD Concept 48	Focal post’s variables
32	29	Deviation_subjectivity	Leading variables
33	29	SVD Concept 10	Focal post’s variables
34	29	Mean_polarity	Leading variables
35	29	SVD Concept 6	Focal post’s variables
36	29	SVD Concept 81	Focal post’s variables
37	29	SVD Concept 34	Focal post’s variables
38	28	SVD Concept 78	Focal post’s variables
39	28	Number_characters	Focal post’s variables
40	28	SVD Concept 13	Focal post’s variables
41	28	SVD Concept 83	Focal post’s variables
42	28	SVD Concept 9	Focal post’s variables
43	28	SVD Concept 3	Focal post’s variables
44	28	SVD Concept 25	Focal post’s variables
45	28	SVD Concept 63	Focal post’s variables
46	28	SVD Concept 53	Focal post’s variables
47	28	SD_time_likes	Leading variables
48	28	SVD Concept 18	Focal post’s variables
49	28	SVD Concept 7	Focal post’s variables
50	28	Nbr_comments	Lagging variables

References

[1] A. Abbasi, H. Chen, A. Salem, Sentiment analysis in multiple languages: feature selection for opinion classification in Web forums, ACM Transactions on Information Systems 26 (3) (2008) 12:1–12:34.

[2] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, Sentiment Analysis of Twitter Data, Proceedings of the Workshop on Languages in Social Media, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 30–38.

[3] E. Alpaydin, Combined 5 times 2 cv F Test for comparing supervised classification learning algorithms, Neural Computation 11 (8) (1999) 1885–1892.

[4] X. Bai, R. Padman, E. Airoldi, Sentiment Extraction from Unstructured Text using Tabu Search-Enhanced Markov Blanket, Technical report, Carnegie Mellon University, School of Computer Science, Technical Report CMU-IS-RI-04-127, 2004.

[5] E. Bakshy, D. Eckles, R. Yan, I. Rosenn, Social Influence in Social Advertising: Evidence from Field Experiments, Proceedings of the 13th ACM Conference on Electronic Commerce, ACM, New York, NY, USA, 2012, pp. 146–161.

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

[7] M. Ballings, D. Van den Poel, interpretR: Binary Classifier and Regression Model Interpretation Functions, Jun. 2015.

[8] M. Ballings, D. Van den Poel, N. Hespeels, R. Gryp, Evaluating multiple classifiers for stock price direction prediction, Expert Systems with Applications 42 (20) (Jun. 2015) 7046–7056.

[9] L. Barbosa, J. Feng, Robust Sentiment Detection on Twitter from Biased and Noisy Data, Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 36–44.

[10] M.E. Basiri, N. Ghasem-Aghaee, A.R. Naghsh-Nilchi, Exploiting reviewers comment histories for sentiment analysis, Journal of Information Science 40 (3) (2014) 313–328.

[11] E. Bast, C. Kuzey, D. Delen, Analyzing initial public offerings' short-term performance using decision trees and SVMs, Decision Support Systems 73 (2015) 15–27.

[12] R.F. Baumeister, E. Bratslavsky, C. Finkenauer, K.D. Vohs, Bad is stronger than good, Review of General Psychology 5 (4) (2001) 323–370.

[13] S. Ben Hamouda, J. El Akaichi, Social networks text mining for sentiment classification: the case of Facebook statuses updates in the Arabic Spring Era, International Journal of Application or Innovation in Engineering and Management 2 (5) (2013) 470–478.

[14] A. Ben-Hur, J. Weston, A User Guide to Support Vector Machines, in: O. Carugo, F. Eisenhaber (Eds.), Data Mining Techniques for the Life Sciences, 609, Humana Press, Totowa, NJ, 2010, pp. 223–239.

[15] B. Blamey, T. Crick, G. Oatley, R U :-) or :-(? Character- vs. Word-Gram Feature Selection for Sentiment Classification of OSN Corpora, in: M. Bramer, M. Petridis (Eds.), Research and Development in Intelligent Systems XXIX, Springer London. 2012, pp. 207–212.

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

[17] J. Bollen, A. Pepe, H. Mao, Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena, arXiv:0911.1583 [cs] (2009)

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

[19] Q. Cao, W. Duan, Q. Gan, Exploring determinants of voting for the helpfulness of online user reviews: A text mining approach, Decision Support Systems 50 (2) (2011) 511–521.

[20] P. Chaovalit, L. Zhou, Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches, Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 2005. HICSS '05, 2005. pp. 112c-112c.

[21] N.A. Christakis, J.H. Fowler, Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives - How Your Friends' Friends' Friends Affect Everything You Feel, Think, and Do, reprint edition ed., Back Bay Books, New York, NY u.a., 2011.

[22] A.U. CLiPS, Pattern - Web mining module for Python, 2014, URL https://github.com/clips/pattern.

[23] K. Coussement, D. Van den Poel, Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques, Expert Sytems with Applications 34 (1) (2008) 313–327.

[24] N.F.F. da Silva, E.R. Hruschka, E.R. Hruschka, Jr, Tweet sentiment analysis with classifier ensembles, Decision Support Systems 66 (2014) 170–179.

[25] L. Dang-Xuan, S. Stieglitz, Impact and Diffusion of Sentiment in Political Communication An Empirical Analysis of Political Weblogs, Sixth International AAAI Conference on Weblogs and Social Media, 2012. pp. 1–4.

[26] K. Dave, S. Lawrence, D.M. Pennock, Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews, Proceedings of the 12th International Conference on World Wide Web, ACM, New York, NY, USA, 2003, pp. 519–528.

[27] D. Davidov, O. Tsur, A. Rappoport, Enhanced Sentiment Learning Using Twitter Hashtags and Smileys, Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 241–249.

[28] L. de Vries, S. Gensler, P.S.H. Leeflang, Popularity of brand posts on brand fan pages: an investigation of the effects of social media marketing, Journal of Interactive Marketing 26 (2) (2012) 83–91.

[29] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, Journal of the American Society for Information Science 41 (6) (1990) 391–407.

[30] J. Demšar, Statistical comparisons of classifiers over multiple data sets, The Journal of Machine Learning Research 7 (2006) 1–30.

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

[32] E. Diener, Subjective Well-Being and Personality, in: D.F. Barone, M. Hersen, V.B.V. Hasselt (Eds.), Advanced Personality. The Plenum Series in Social/Clinical Psychology, Springer, US, 1998, pp. 311–334.

[33] T.G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Computation 10 (7) (1998) 1895–1923.

[34] S. Dudoit, J. Fridlyand, T.P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, Journal of the American Statistical Association 97 (457) (2002) 77–87.

[35] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research 15 (2014)31333181.

[36] E. Fersini, E. Messina, F.A. Pozzi, Sentiment analysis: Bayesian ensemble learning, Decision Support Systems 68 (2014) 26–38.

[37] A.L. Forest, J.V. Wood, When social networking is not working individuals with low self-esteem recognize but do not reap the benefits of self-disclosure on Facebook, Psychological Science (2012)0956797611429709.

[38] W. Frakes, R. Baeza-Yates, Information Retrieval: Data Structures and Algorithms, Prentice Hall PTR. 1992.

[39] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, Journal of the American Statistical Association 32 (200) (1937) 675–701.

[40] M. Gamon, Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis, Proceedings of the 20th International Conference on Computational Linguistics, No. 841, Association for Computational Linguistics. Vol. No. 841 of COLING ′04, Stroudsburg, PA, USA, 2004, pp. 1–7.

[41] A. Go, R. Bhayani, L. Huang, Twitter sentiment classification using distant supervision, Technical report, CS224N Project Report, Stanford, 2009.

[42] I. Habernal, T. Ptek, J. Steinberger, Supervised sentiment analysis in Czech social media, Information Processing & Management 50 (5) (2014) 693–707.

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

[44] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer Series in Statistics Springer New York, New York, NY, 2009.

[45] E. Hatfield, J.T. Cacioppo, R.L. Rapson, Emotional contagion, Studies in emotion and social interaction. vii. Editions de la Maison des Sciences de l'Homme, Paris, France, 1994.

[46] V. Hatzivassiloglou, J.M. Wiebe, Effects of Adjective Orientation and Gradability on Sentence Subjectivity, Proceedings of the 18th Conference on Computational Linguistics - Volume 1, Association for Computational Linguistics, Stroudsburg, PA, USA, 2000, pp. 299–305.

[47] J.F. Helliwell, R.D. Putnam, The social context of well-being, Philosophical Transactions of the Royal Society B: Biological Sciences 359 (1449) (2004) 1435–1446.

[48] C.-W. Hsu, C.-C. Chang, C.-J. Lin, A practical guide to support vector classification, Tech. rep., Department of Computer Science, National Taiwan University. 2003,

[49] D. Huffaker, Dimensions of Leadership and Social Influence in Online Communities, Human Communication Research 36 (4) (2010) 593–617.

[50] A.M. Kaplan, M. Haenlein, Users of the world, unite! The challenges and opportunities of Social Media, Business Horizons 53 (1) (2010) 59–68.

[51] M. Kosinski, Y. Bachrach, P. Kohli, D. Stillwell, T. Graepel, Manifestations of user personality in website choice and behaviour on online social networks, Machine Learning 95 (3) (2013) 357–380.

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

[53] E. Kouloumpis, T. Wilson, J. Moore, Twitter Sentiment Analysis: The Good the Bad and the OMG!, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011. pp. 538–541.

[54] W. Kraaij, R. Pohlmann, Porters stemming algorithm for Dutch, Informatiewetenschap 1994: Wetenschappelijke bijdragen aan de derde STINFON Conferentie, 1994. pp. 167–180.

[55] A.D. Kramer, An Unobtrusive Behavioral Model of "Gross National Happiness", Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, 2010, pp. 287–290.

[56] A. Kumar, T.M. Sebastian, Sentiment analysis on Twitter, IJCSI International Journal of Computer Science Issues 9 (4(3)) (2012) 372–378.

[57] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1) (1977) 159.

[58] N. Li, D.D. Wu, Using text mining and sentiment analysis for online forums hotspot detection and forecast, Decision Support Systems 48 (2) (2010) 354–368.

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

[60] H.L. Lin Qiu, Putting their best foot forward: emotional disclosure on Facebook, Cyberpsychology, behavior and social networking 15 (10) (2012) 569–572.

[61] B. Liu, Synthesis Lectures on Human Language Technologies, Sentiment Analysis and Opinion Mining 5, Morgan & Claypool Publishers. 2012, pp. 1–167.

[62] E. Martínez-Cámara, M.T. Martín-Valdivia, L.A. Ure na López, A. Montejo-Ráez, Sentiment analysis in Twitter, Natural Language Engineering 20 (1) (2014) 1–28.

[63] S. Matsumoto, H. Takamura, M. Okumura, Sentiment Classification Using Word Sub-sequences and Dependency Sub-trees, in: T.B. Ho, D. Cheung, H. Liu (Eds.), Advances in Knowledge Discovery and Data Mining. No. 3518 in Lecture Notes in Computer Science, Springer, Berlin Heidelberg, 2005, pp. 301–311.

[64] B. McInnes, Supervised and Knowledge-based Methods for Disambiguating Terms in Biomedical Text Using the Umls and Metamap, University of Minnesota, Minneapolis, MN, USA, 2009. (Ph.D. thesis)

[65] P. Melville, W. Gryc, R.D. Lawrence, Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification, Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2009, pp. 1275–1284.

[66] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, e1071: Misc Functions of the Department of Statistics (e1071), TU Wien, 2014. C++-code), C.-C. C. l., C++-code), C.-C. L. l., Sep. 2014.

[67] S.M. Mohammad, S. Kiritchenko, Using hashtags to capture fine emotion categories from Tweets, Computational Intelligence 31 (2) (2015) 301–326.

[68] T. Mullen, N. Collier, Sentiment analysis using support vector machines with diverse information sources, Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2004. pp. 412–418.

[69] F. Neri, C. Aliprandi, F. Capeci, M. Cuadros, T. By, Sentiment Analysis on Social Media, Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), IEEE Computer Society, Washington, DC, USA, 2012, pp. 919–926.

[70] M.W. Newman, D. Lauterbach, S.A. Munson, P. Resnick, M.E. Morris, It's Not That I Don'T Have Problems, I'M Just Not Putting Them on Facebook: Challenges and Opportunities in Using Online Social Networks for Health, Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, ACM, New York, NY, USA, 2011, pp. 341–350.

[71] A. Ortigosa, R.M. Carro, J.I. Quiroga, Predicting user personality by mining social interactions in Facebook, Journal of Computer and System Sciences 80 (1) (2014) 57–71.

[72] A. Ortigosa, J.M. Martn, R.M. Carro, Sentiment analysis in Facebook and its application to e-learning, Computers in Human Behavior 31 (2014) 527–541.

[73] A. Pak, P. Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining, Proceedings of LREC,2010, 2010. pp. 1320–1326.

[74] B. Pang, L. Lee, Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval 2. No. 2 in 2., Now Publishers Inc. 2008, pp. 1–135.

[75] B. Pang, L. Lee, S. Vaithyanathan, Thumbs Up?: Sentiment Classification Using Machine Learning Techniques, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, pp. 79–86.

[76] J.W. Pennebaker, M.R. Mehl, K.G. Niederhoffer, Psychological aspects of natural language use: our words, our selves, Annual Review of Psychology 54 (1) (2003) 547–577.

[77] M. Porter, An algorithm for suffix stripping, Program 14 (3) (1980) 130–137.

[78] R. Prabowo, M. Thelwall, Sentiment analysis: a combined approach, Journal of Informetrics 3 (2) (2009) 143–157.

[79] D. Quercia, J. Ellis, L. Capra, J. Crowcroft, Tracking "Gross Community Happiness" from Tweets, Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, ACM, New York, NY, USA, 2012, pp. 965–968.

[80] J. Read, Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification, Proceedings of the ACL Student Research Workshop, Association for Computational Linguistics, Stroudsburg, PA, USA, 2005, pp. 43–48.

[81] E. Riloff, S. Patwardhan, J. Wiebe, Feature Subsumption for Opinion Analysis, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 2006, pp. 440–448.

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

[83] H.A. Schwartz, J.C. Eichstaedt, M.L. Kern, L. Dziurzynski, S.M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M.E.P. Seligman, L.H. Ungar, Personality, gender, and age in the language of social media: the open-vocabulary approach, PLoS ONE 8 (9). (2013)e73791.

[84] M. Settanni, D. Marengo, Sharing feelings online: studying emotional well-being via automated text analysis of Facebook posts, Personality and Social Psychology (2015) 1045.

[85] I. Smeureanu, C. Bucur, Applying supervised opinion mining techniques on online user reviews, Informatica Economica 16 (2) (2012) 81–91.

[86] S.M. Smith, R.E. Petty, Message framing and persuasion: A message processing analysis, Personality and Social Psychology Bulletin 22 (3) (1996) 257–268.

[87] S. Stieglitz, L. Dang-Xuan, Impact and diffusion of sentiment in public communication on Facebook, ECIS 2012 Proceedings (2012)

[88] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Computational Linguistics 37 (2) (2011) 267–307.

[89] A. Tamilselvi, M. ParveenTaj, Sentiment analysis of micro blogs using opinion mining classification algorithm, International Journal of Science and Research (IJSR) 2 (10) (2013) 196–202.

[90] S. Tan, Y. Wang, X. Cheng, Combining Learn-based and Lexicon-based Techniques for Sentiment Detection Without Using Labeled Examples, Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2008, pp. 743–744.

[91] C. Troussas, M. Virvou, K. Junshean Espinosa, K. Llaguno, J. Caro, Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning, 2013 Fourth International Conference on Information, Intelligence, Systems and Applications (IISA), 2013. pp. 1–6.

[92] A. Tumasjan, T.O. Springer, P.G. Sandner, I.M. Welpe, Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 2010. pp. 178–185.

[93] P.D. Turney, Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, pp. 417–424.

[94] S. Wang, C.D. Manning, Baselines and Bigrams: Simple, Good Sentiment and Topic Classification, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 90–94.

[95] Wikipedia, List of emoticons, 2015, URL http://en.wikipedia.org/w/index.php?title=List_of_emoticons&oldid=654618502.

[96] H. Yu, V. Hatzivassiloglou, Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences, Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 2003, pp. 129–136.

[97] Y. Yu, X. Wang, World Cup 2014 in the Twitter world: a big data analysis of sentiments in US sports fans' tweets, Computers in Human Behavior 48 (2015) 392–400.

[98] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, B. Liu, Combining lexicon-based and learning-based methods for Twitter sentiment analysis, HP Laboratories. 2011.

Matthijs Meire is a PhD student at Ghent University. He has received his B.S. in Business Engineering and M.S. in Business Engineering: Marketing Engineering/Data Analytics from Ghent University. His research interests are text mining, social media analytics and customer analytics.

Michel Ballings (PhD) is Assistant Professor of Business Analytics at The University of Tennessee (Knoxville). He teaches Data Mining and Customer Analytics. His research interests are in social media analytics, customer analytics, and machine learning. He has co-authored several peer-reviewed publications in journals such as European Journal of Operational Research, Omega, Decision Support Systems, and Expert Systems with Applications.

Dirk Van den Poel (PhD) is a Professor of Business Analytics/Big Data at Ghent University, Belgium. He teaches courses such as Statistical Computing, Big Data, Analytical Customer Relationship Management, Advanced Predictive Analytics, Predictive and Prescriptive Analytics. He cofounded the advanced Master of Science in Marketing Analysis, the first (predictive) analytics master program in the world as well as the Master of Science in Statistical Data Analysis and the Master of Science in Business Engineering/Data Analytics. His major research interests are in the field of analytical CRM (Customer Relationship Management): customer acquisition, churn, upsell/cross-sell, and win-back modeling. His methodological interests include ensemble classification methods and big data analytics. He has co-authored 70+ international peer-reviewed (ISI-indexed) publications in journals such as Journal of Applied Econometrics, Applied Geography, European Journal of Operational Research, and Decision Support Systems.

Please cite this article as: M. Meire, et al., The added value of auxiliary data in sentiment analysis of Facebook posts, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.06.013