Predicting Social Media Duration Classes:

A Machine Learning Approach

Joeri van de Rijdt

June 25, 2020

u1266030 / 460157

Tilburg University

Tilburg School of Humanities and Digital Sciences

MSc. Thesis Data Science & Society 2020

Department of Cognitive Science and Artificial Intelligence

Supervisor: dr. D. Hendrickson

Abstract

The purpose of this study is to predict the correct social media usage class of individuals. The basis for these classes is the duration that individuals spend on social media platforms. Different machine learning algorithms are utilized to address this problem. The data at hand consists of phone tracking data and mood survey data. The main question to be answered is to what extent machine learning algorithms can predict these social media classes, using a combination of the two mentioned data types. This problem as well as the combination of phone and mood features have not been studied in literature before. The metrics to evaluate the performance of the models are accuracy and recall. As it turns out, all models outperform their benchmark when it comes to accuracy. With regard to recall, only some of the models outperform their benchmark. Recall is a more important metric than accuracy to predict problematic social media usage. Overall, the predictive value of the machine learning algorithms is not large enough to have an impact on businesses. There are several opportunities for future research.

Table of Contents

1. Introduction

The consequences of the globalization of mobile phones

Initially, the globalization of mobile phones seemed beneficial. It brings people closer together and it increases GDP per capita growth (Nguwi & Loh, 2017; Argawal & Lu, 2019). Then, this turned around and the negative consequences have now been pointed out frequently. Examples of these negative consequences are lower mood and self-esteem and even depression (Berry, Emsley, Lobban & Bucci, 2018). An important factor in this switch is the use of social media applications. Individuals use social media platforms when their mood is low and using social media platforms, in turn, has a detrimental impact on mood, possibly even leading to depression (Robinson et al., 2019). Therefore, preventing problematic social media usage is of societal importance.

Social media class

The goal of this study was to predict the correct social media class (hereinafter "SMC") that individuals belong to. These classes are constructed based on the time that individuals spend on social media platforms. When a participant belongs to the fraction of individuals that spends the largest amount of time on social media applications, this participant belongs to the highest SMC for a certain time interval. Might the SMCs be divided over four classes, then this participant belongs to class '4'. The goal of this study is to assess the value of different machine learning algorithms in predicting this SMC feature, using a variety of other features.

Academic relevance

Several machine learning algorithms were used and assessed on their predictive value. The algorithms were used to classify instances in both binary and multiclass models. Two types of features are used: phone features that are constructed from phone tracking data and mood features that are constructed from mood survey data. This problem has not been investigated in existing literature, yet a variety of aspects regarding this problem have been addressed. It is confirmed that internet addiction is a result of internal stress (Kuang-Tsan & Fu-Yuan, 2017). Furthermore, the highest quartile of social media users are significantly more likely to become depressed (Lin et al, 2016). This makes the subject an important societal issue. Even though the importance is there, no study has yet tried to predict social media classes. In addition, an elaboration of the differences between phone and mood features is lacking as well.

Research questions

In order to study this topic, several research questions will be answered. First, the problem statement will be addressed:

- To what extent can machine learning models predict the correct and/or most problematic SMC, using a combination of phone and mood features?

To answer the problem statement, several sub-questions need to be answered as well:

- Which factors play a role in predicting SMCs?

Apart from the phone and mood features at hand, several other factors might influence social media usage. Existing literature will be critically reviewed to find other factors that explain social media usage.

- To what extent does the additional value of using features from mood survey data outweigh the additional costs of acquiring this data?

Mood survey data is more expensive than phone tracking data (Singh & Long, 2018; Pratap et al., 2018). Yet, knowing several mood characteristics for the participants in a social media study like this might generally result in models with significantly better performances. Therefore, it is useful to assess how the machine learning models perform when features constructed from mood survey are added.

- To what extent can machine learning algorithms outperform their benchmark in predicting SMC?

Two metrics that are important regarding this topic will be evaluated: accuracy and recall. The goal, when building these algorithms, is to outperform a benchmark. Since the classes are constructed using quartile splitting, the benchmark to outperform is always (1 / amount of classes). In this study, a binary classification problem is tackled using two classes, whereas a multiclass classification problem incorporates a total of four classes. Therefore, the benchmark is 50% for binary classifiers and 25% for multiclass problems.

Main findings

The findings are predominantly optimistic in terms of accuracy. All models have a higher accuracy than their benchmark. In terms of recall, the models do not have sufficient predictive power to predict the highest SMC. Additionally, there is a slight improvement for both accuracy and recall when mood features are incorporated next to phone features. When either kind of features are used, the phone features score higher in general.

The following section gives background information. Section 3 gives an explanation of the algorithms that are used in this study. The fourth section is a narration of the experimental procedure that is followed. Section 5 shows the results and a discussion of these results is given in section 6. The last section concludes.

2. Background

Effects of mobile phones

Mobile phones connect businesses and people on a global level. The advantages and profits that mobile telephones bring are undeniable: It increases GDP per capita growth, brings people closer to one other and entertains people (Nguwi & Loh, 2017). Moreover, mobile phones provide ways to stay in touch with others and allow people to multitask. For instance, by listening to music and texting, while walking at the same time. Also, a lot of people do not experience their heavy use as a problem and feel that they are meaningfully involved in an online community (Argawal & Lu, 2019).

On the other side of the spectrum are the negative consequences of mobile phone globalization. Almost all of the positive consequences that are mentioned above come with certain disadvantages. To point out some examples: it is difficult to express your actual intentions or emotions through text messages, multitasking is also distracting or counterproductive and overuse is actually unhealthy because it is easy to waste your life (Argawal & Lu, 2019).

Social media: definitions and why it is used

With the globalization of the mobile phone also came the rise of social media platforms. There seems to be no consensus about the definition of social media in the literature. To point out some examples, Wood, Bukowski and Lis (2016) describe it as several technologies that make immediate communication possible, also allowing for status updates and social networking among individuals. Another definition is pointed out by Theem, Dailey, Pierce and Biffl (2016). They explain it as a way for individuals to maintain current relationships, create new connections, create and share content, and, in some degree, make their own social networks observable to others. Despite the debate on how exactly the definition of social media should be formulated, there seems to be more consensus about what goals social media is used for. Individuals use social media primarily for entertainment, social interaction, to seek information and to pass time (Whiting & Williams, 2013; Griffioen, Van Rooij, Lichtwarck-Aschoff & Granic, 2020).

Effects of social media

Apart from these goals, the consequences of social media usage are detrimental. A well-established consequence of social media is that it results in a decrease in mood. Posting about feelings and venting on social media, as well as perceptions of low social rank when using social media, both predict low mood and self-esteem (Berry, Emsley, Lobban & Bucci, 2018). This lower mood can then result in an addiction. More specifically, poor emotion regulation and emotional dysregulation are important factors in the development of addiction in general (Nguwi & Loh, 2017). In reducing emotional destress, an individual may direct his/her attention to seek immediate pleasure and relief. Problematic phone use may also serve as a short-term strategy to deal with unpleasant emotions or moods. Authors Kuang-Tsan and Fu-Yuan (2017) confirm this by stating that specifically internet addiction is a result of individuals being under internal stress. It then becomes a method to release tension and daily pain, as supported by empirical studies.

Another possible consequence of social media usage is depression. A nationally-representative study among United States young adults points out that participants in the highest social media usage quartile have significantly increased odds of depression (Lin et al., 2016). This study will take the results from Lin et al. (2016) into account when predicting social media usage classes, especially the highest and most problematic class. Since the highest social media usage quartile is the most harmful according to the literature, flagging as much as the highest-quartile social media uses correctly is of interest. Another study complements this by showing that millennials who actively use social media applications are more likely to meet the criteria for Major Depressive Disorder (MDD) (Robinson et al., 2019). Altogether, the negative implications of social media usage are undeniable. It is of societal importance to predict and prevent problematic use.

Apart from the strong claims on the negative consequences of social media usage, there is also evidence that this connection does not exist. Orben, Dienlin and Przybylski (2019) use a large-scale, representative dataset from which they investigate the effect of social media usage on the life satisfaction of adolescents. The authors find that social media usage is not a strong predictor for life satisfaction. The relationship between these variables is weak 'at best'. Also, this relationship is more nuanced and some aspects of the relationship require more exploration, according to the authors. Similar statements are made by El-Badawy & Hashem (2015), who strongly claim that there is no relationship between social media usage and academic performance of students.

The relationship between mood and social media

An important topic in this study is the relationship between mood and social media. Most studies aim to predict certain mood characteristics from social media usage (Argawal & Lu, 2019; Berry, Emsley, Lobban & Bucci, 2018; Kuang-Tsan & Fu-Yuan, 2017; Lin et al., 2016; Robinson et al, 2019; Orben, Dienlin & Przybylski, 2019; El-Badawy & Hashem, 2015; Pratap et al., 2018; Singh & Long, 2018). The newness in the current lies in switching the direction of the variables. Only Nguwi and Loh (2017) try to predict problematic phone usage from emotion dysregulation, leaving social media disregarded. This is important to keep in mind, as the current study aims to predict social media usage from phone and mood data, yet almost all of existing literature focuses on predicting mood from phone features.

Predicting social media classes: mood features

One way to predict problematic social media usage is by utilizing machine learning models. The current study will assess several machine learning models that aim to predict the correct social media usage class of individuals. The models will predict social media classes utilizing two types of features: mood features using surveys and phone usage features using phone trackers. Mood features are of interest due to the aforementioned relationship between mood and social media usage. The idea of assessing individuals' mood through surveys originates from the healthcare industry. Assessing the mental well-being of individuals accurately is important for building well-functioning societies (Singh & Long, 2018). Traditional methods for assessing mental health and well-being are surveys that are used by healthcare professionals. Such surveys are labour-intensive, costly, and typically intermittent.

With the rise of the mobile phone, a number of recent efforts have tried to utilize mobile phones to use surveys that understand user state (Singh & Long, 2018). For instance, Servia-Rodriguez et al. (2017) discuss a large-scale study aimed at obtaining user state multiple times a day over time. Such mood-assessing surveys are also what constructed the mood dataset in this study. A mobile phone application called Ethica is used to send surveys four times a day to the participants to assess their mood. Ethica is a platform to quantitatively measure human behaviour using smartphones and big data (www.ethicadata.com). A variety of emotions are assessed by the surveys, namely: anxious, bored, gloomy, calm, stressed, content, cheerful, tired, energetic, upset and envious. The participants could then answer on a six-point Likert scale per question ranging from 0 to 5 whether they agreed that their feelings correspond to the emotions.

Predicting social media classes: phone features

Phone usage features are the other type of features that are used in the current study. With the mentioned growth in mobile sensing, health informatics and data science, the problem of preventing problematic social media usage calls for a method that uses mobile phone meta-data and data analytics to automatically infer an individual's mental health (Singh & Long, 2018). This is exactly the case in current literature, where phone usage is often utilized to predict the mood of individuals. An example is the study of Pratap et al. (2018), where the authors try to predict mood on a daily basis. The results of this paper are favourable in terms of predictiveness.

Smartphones provide a low-cost and efficient means to collect population level data. Several small studies have shown promise in predicting mood from usage data. On the other hand, Pratap et al. (2018) claim that passive smartphone data is not suited for predicting daily mood, due to the high degree of intra- and interindividual variation in phone usage patterns and daily mood ratings. This study will assess these findings from another perspective, namely by switching the independent and dependent variables. Instead of predicting mood from phone usage, this study will predict social media phone usage from mood and the use of other applications. To construct the phone usage data, again, the Ethica application is used. The begin and ending time of every application opening are registered and eventually stored in one big dataset, containing the data for all participants. An average of 4,732 application openings are registered per participant over the horizon of this study. Phone tracking data is cheaper and easier to obtain than mood survey data (Singh & Long, 2018; Pratap et al., 2018).

Adolescents as study object

The sample in this study consists of university students. This kind of individuals has been investigated before to explain the relationship between academic stress (being a part of mood) and mobile phone addiction (Kuang-Tsan & Fu-Yuan, 2017). Yet, this study regards only stress as mood feature and does not specify phone usage to the social apps. Lin et al. (2016) also used young adults in their research and found that individuals in the highest social media usage quartile have a higher probability of becoming depressed. However, the authors do not aim to predict the social media class, nor do they use mood features as predictors.

For the same kind of individuals, university students, checking social media accounts leads to academical disengagement in class (Gupta & Irwin, 2016). A classroom test which allowed mobile phone usage for one group and disallowed cell phones for another group showed significant worse results in test performance for the phone users (Lee et al., 2017). Academic stress and test performance are concepts that belong to university students or young adults in general. However, the current literature does not cover the generalizability of these influences of social media usage over other age categories in different settings yet. This phenomenon will be discussed more extensively in the discussion section.

Utility changes and gender differences

Apart from the research field being generalizable over other categories of individuals, there are several other aspects to keep in mind as well as. First of all, there is an indication that the utility of certain social media platforms evolves (Idemudia, Raisinghani & Samual-Ojo, 2016). This means that some social media platforms become less interesting over time and might not result in negative mental consequences anymore.

Also, there is some evidence that the effects of social media are more negative for females than for males (Wood, Bukowski & Lis, 2016). Males may benefit from social media use by building social skills, where females experience more negative consequences, like lower self-esteem. This claim is underwritten by Orben, Dienlin and Przybylski (2019), who also state that the relationship between social media usage and life satisfaction is significantly more negative for females than for males. Note that the participants in the current study are anonymized and their gender is unknown. The utility evolvement of social media applications and the differences between genders are interesting aspects that currently rise in literature with the still increasing degree of social media usage.

Machine learning models

In order to find out which machine learning models are most appropriate for this classification task, several studies have been consulted. Unfortunately, existing literature only consists of some studies that study aspects regarding the current study. For instance, Singh and Long (2018) try to automatically assess the mental health of individuals, also using phone usage features. In this classification task, they use logistic regression, Naive Bayes, random forest and ZeroR classifiers and state that Naive Bayes and Random Forest perform best in their investigation. Nguwi and Loh (2017) switch the topic of interest by examining whether problematic phone usage can be predicted by impulsivity and emotion dysregulation. They use the same machine learning algorithms, adding a multilayer perceptron. Van Zoonen and Van der Meer (2016) categorize social media content using a Linear Support Vector Machine, Naive Bayes and Logistic Regression. The authors state that the latter underperforms.

Since the support vector machine, Naive Bayes and random forest algorithms seem to perform well in comparable tasks according to the literature, these algorithms will be used in the current study as well. Since the logistic regression algorithm underperforms other algorithms in different studies, this algorithm will not be considered.

Constructing SMC and evaluating results

Singh and Long (2018) performed median split in their paper on mobile phone surveys to understand user state. In a set of numbers that are arranged in order, the median is the middle number. As a result, half of the observations are lower than the median and half of the observations are higher than the median. Hence, using median split while working on binary classification problems ensures that the classes are balanced.

Constructing social media duration quartiles is performed by Van Zoonen and Van der Meer (2016). The authors explicitly state that this is preferable over constructing quartiles based on social media usage frequency. Using quartiles to construct the outcome variable in a four-class classification problem also ensures that the classes are balanced, as a quarter of the total number of observations are assigned to every class. In the current study, the SMCs are constructed based on duration, using median split for the binary classifiers and quartiles for the multi-class classifiers. This is in accordance with Singh and Long (2018) and Van Zoonen and Van der Meer (2016).

Within the machine learning models, a distinction between the cheap phone use data and the expensive mood assessment data will be made. The models will be evaluated on both accuracy and recall. Accuracy will show many of the classes are predicted correctly, whereas recall, in this case, will reveal the amount of the highest quartile social media usages that are also identified as such. The latter is interesting from both a societal and business perspective. Some business opportunities have already been addressed.

Business opportunities to overcome negative consequences

Looking at the proliferation of social media, identifying the mechanisms and direction of the association with depression is critical for interventions that address problematic social media usage and depression (Lin et al., 2016). The current research plays a role in identifying the mechanisms and directions of social media usage and mood. Such interventions are particularly interesting from a business perspective. There are several ways to address individuals with problematic social media usage. For example, MoodMission is an application that is claimed to improve individuals' feelings. Such applications might help overcoming the issue of problematic social application usage and the negative mood that it comes with (Bakker & Rickard, 2019). Another way to overcome this issue is interventions aimed at permanently improving people's subjective well-being through big data. This is an idea that Luhmann (2017) proposed. By means of example, self-tracking apps with built-in interventions such as feedback on one's self-reported mood, or specific advice on how to improve one's mood, can provide a solution for several mental problems.

3. Methods

This section gives an explanation of the mathematical algorithms that are utilized in this study. Note that for each algorithm, the reason of use and applicability in the current study is described in this section, though the way they are implemented and the results of these implementations can be found in the experimental setup and results sections, respectively. The order of the algorithms in this section is the same as the order in which they are used in this study.

3.1 Principal Component Analysis

Principal Component Analysis (PCA) was invoked around 120 years ago (Pearson, 1901). PCA involves a dimensionality reduction process. The initial data is reconstructed in new features by performing eigenvalue decomposition. A PCA plot can then show to what extent certain features contribute to the variability in the data.

I performed PCA to find the most informative features in the data. I did so by using the *prcomp* function from the *stats* package in R, which allowed me to incorporate all features that I constructed. The result is a graph that shows which features contribute most to the first and second PC. A more detailed outcome will be described in the experimental setup section.

3.2 Naive Bayes

Naive Bayes is a statistical machine learning algorithm that uses conditional probabilities to solve problems (Hand & Yu, 2001). The application of Naive Bayes in the current study was straightforward as the *naiveBayes* function in R allows users to specify the feature of interest and the features you want to utilize in predicting this feature of interest. In other words, I could specify SMC as the dependent feature and all other remaining features as the predictors. Then, after training the model, the SMC feature is predicted for several instances in the test set. The *confusionMatrix* function is used to assess the accuracy and recall of the model. A more detailed description of the implementation of Naive Bayes in this study can be found in the experimental setup section.

3.3 Random Forest

The Random Forest algorithm is a computational mechanism that consists of multiple decision trees (Breiman, 2001). A decision tree is an algorithm that splits the dataset by features and ultimately votes for a certain class. A random forest consists of multiple decision trees, each voting for a certain outcome.

In this study, the trees are all trained to vote for either of the SMCs and the mode of these votes is what the complete random forest predicts. That is how the model is trained using the training set. The test set is used to let the trees vote for a certain SMC given all other feature values. Again, a confusion matrix is constructed to assess the performance for every version of the model. Please see the experimental setup for more details on how the random forest algorithm is implemented in R.

3.4 Support Vector Machine

Another machine learning algorithm applied in this study is a Support Vector Machine (SVM). This algorithm works in such a way that for certain input data, a decision boundary is constructed to decide whether inputs fall in one class or another. The goal is to assign new instances to the correct category (Evgeniou & Pontil, 2001).

In this study, a non-linear SVM that uses a Radial Basis Function kernel is utilized. This means that the kernel trick is also performed and the SVM space is a multidimensional space where classification boundaries are curved lines. Further details on how to implement the SVM algorithm in R can be found in the experimental setup.

4. Experimental Setup

4.1 Dataset description

Mood survey data

The first dataset contains mood survey data. It consists of 16,016 observations of 35 variables. Various variables are disregarded as they do contribute to this study. The variables that have a contribution are explained. The Ethica application, which also tracked the data in the phone usage dataset, is used to send four surveys per day to the participants. A total of 149 participants are distinguished in this dataset. The dataset shows the non-identifiable *user ID*, the time the survey was sent and the time at which the survey was completed. Also, the dataset contains answers on twelve questions regarding the participants' mood and a daytime window variable. The daytime window variable indicates in which part of day the observation is made. There are four daytime windows per day, all translating to a certain time interval. An important note is that after re-constructing the daytime window variable into a more balanced one, several observations are lost. This is due to for instance observations which occur at night time that are not part of the newly constructed variable. The questions regarding mood are based on a six-point Likert scale that ranges from 0 to 5. The twelve kinds of mood in the survey are: *anxious*, *bored*, *gloomy*, *calm*, *stressed*, *content*, *cheerful*, *tired*, *energetic*, *upset*, *envious* and *inferior*. The dataset is ordered per *user ID*, so one can easily find all surveys that belong to one single participant.

Categorical data

The second dataset categorizes the applications. For 1,748 different applications, the dataset indicates to which category a certain application belongs. For instance, WhatsApp is categorized as a communication application. It also contains *better category hybrid* and *better category* as variables. To remain with the WhatsApp example, *better category hybrid* translates to 'WhatsApp Messenger' and *better category* translates to 'Instant Messages'. Furthermore, the *count* of every application is stated and the dataset is ordered along this variable in decreasing order. This means that the dataset begins with the most used application (WhatsApp) and ends with the least used application (Color Puzzle Game). The category dataset only contains categories for applications that are used more than hundred times in the accompanying phone usage dataset.

Phone usage data

The third and final dataset that is used in this study is the phone usage dataset. It consists of 586,793 observations of nine variables. The data is gathered using 'Ethica', an application that tracks exactly at what time the participants open and close mobile phone applications. A total of 124 different individuals are tracked and their average amount of application openings during the tracking period is 4,732. The dataset contains the user identification (ID) number, which is a non-identifiable number that belongs to a natural person. The dataset is ordered in chronological order per *user ID*, which means that the data is structured in the exact order the participants open their applications. Other variables are not regarded as they do not have meaning is this study. The first step in the data process is to clean the data.

4.2 Data Cleaning

Mood survey data

Looking at the duration participants spend filling in the survey, it instantaneously stands out that there are many surveys that are either filled in within a minute, or the survey is expired or blocked. I decided to remove all these instances, directly resulting in losing 7,151 observations, which is around 45% of the total amount of mood surveys. There are now 8,865 observations left for analysis. Apart from that, the variables *cheerful*, *tired* and *energetic* have an illegal value in them. This is the result of one participant clicking on more than one answer during one survey. This problem is resolved by deleting this observation and regarding *cheerful, tired* and *energetic* as integers. Then, the response time variable was split up into a ‘day’ feature and a ‘time’ feature using the *lubridate* package in R. The number of surveys per daytime window can be classified in a balanced way, using the following approach:

- Observations between 09:00 and 12:00 are assigned daytime window ‘1’ (2,324)
- Observations between 12:00 and 15:00 are assigned daytime window ‘2’ (2,194)
- Observations between 15:00 and 19:00 are assigned daytime window ‘3’ (2,222)
- Observations between 19:00 and 23:59 are assigned daytime window ‘4’ (2,122)

The number between brackets that comes after the daytime window denotes the number of observations for this daytime window. Only two observations do not fall in the complete timeframe (09:00 – 23:59) and are deleted. The daytime window classes are now balanced.

Next, the mean of every feature is calculated per daytime window per day per participant. This is done using the *dplyr* package in R, combining the *group_by* and the *summarise_all* functions. This finalizes preparing the mood survey data.

Categorical data

The following applications are considered social media: Facebook, Instagram, Twitter, Snapchat, Tumblr, Pinterest and Reddit. Applications like WhatsApp and Facebook Messenger are regarded a communication application and not a social media application. The difference in categorizing these applications lies primarily in the fact that WhatsApp and Facebook Messenger do not sufficiently satisfy the definition of social media as mentioned in the background section. The main difference lies in the fact that using social media applications, individuals can post updates and compare themselves with other people, one of the main characteristics that leads to lower mood.

The next step is categorizing all other applications correctly. There are twelve applications that have no category yet. Seven out of these applications can be categorized by using the same category as all other applications that fall under the same better category hybrid variable. This can be explained with an example. Say the application 'Calculator' would have no *category,* but it does have a *better category hybrid*. All the applications that fall under the same *better category hybrid* are assigned the category 'Tools'. Therefore, Calculator now also is assigned the category 'Tools'. There are four more applications, but for these applications, there is not one single *category* that all the applications under this *better category hybrid* fall under. These five applications are manually assigned to a *category*. This finalizes preparing our category dataset.

Phone usage data

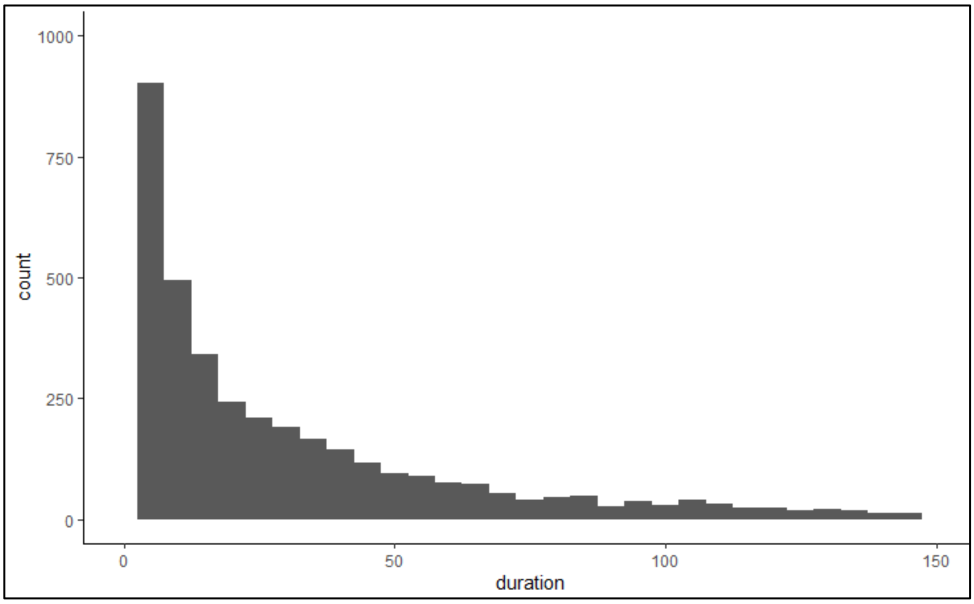
First, the duration of the use of every application is calculated by deducting the begin time from the ending time. The phone usage dataset is then expanded by merging it with the prepared category dataset by application. This action is performed using the *merge* function that is standard in R. If the category for certain applications is missing at this point, this is because the application is used less than hundred times in the original and larger phone usage dataset which the category dataset is based upon. That is why I assign the category 'Low Freq App' to these observations.

Furthermore, there still is a category 'Background Process'. Since applications that are used in your background are not applications that you are actively using at that moment, I decided to delete instances with this category. This results in a deletion of 11,462 observations, still having 575,330 left. The step after data cleaning is feature engineering.

4.3 Feature Engineering

To give a graphical overview of how the duration behaves, I plot it using the *ggplot* function from the *ggplot* package in R. The result is Figure 1.

FIGURE 1: DURATION PLOT IN BINS OF FIVE SECONDS



In Figure 1, the horizontal axis shows the duration in seconds across all observations. As follows from the figure, the vast majority of application uses are less than sixty seconds. The vertical axis shows the count for each of the bins. Since the width of the bins equals five seconds, the first bin shows that around 900 observations concern the use of an application for a duration that lies between zero and five seconds.

Interesting about the duration feature is that its mean is 61.67 seconds, which translates to about one minute, yet the median is 10.74 seconds. This means that there are long application usages that result in a mean that is around six times as large as the median. This is intuitive, since it seems probable that people tend to open most applications for a short amount of time and occasionally watch a movie or another video for a longer time period. Note that these outliers are not found in the graph, since the graph is scaled for readability reasons. The large difference between the mean and the median is a point of issue that comes back later.

Just like in the mood survey data, the time variable is split up in two variables in the phone usage dataset as well. The same daytime hours are used to classify the observations into their daytime window. For this dataset, the result is a loss of 76,042 observations. The lost observations did not fall in any of the earlier explained daytime windows. Constructing these daytime windows enables the eventual data frame to have more instances, which means that more instances are available to train and test the models.

The duration that a certain individual spent in a daytime window using this category is summed for every specific category. This is done by combining the *group_by* and *summarize* functions. This creates a feature for every category that is still in the dataset. By means of example, *Social_Media* is now a feature that illustrates the total duration spent on social media applications per daytime window for every participant.

Next, I merge the datasets by user ID, day and daytime window using the *merge* function. Summarizing the social media feature by using the summary function exposes the quartiles (Q1 = 5.0 seconds, Q2 = 359.0 seconds and Q3 = 1060.0 seconds) of social media usage per daytime window. Constructing the classes by quartiles makes sure that the classes are balanced as perfectly as possible. For multi-class predictions, I make a feature called SMC (Social Media Class) that is constructed as follows:

- Class 1 consists of instances with less than 5 seconds of social media usage.
- Class 2 consists of instances from 5 up until 359 seconds of social media usage.
- Class 3 consists of instances from 359 up until 1060 seconds of social media usage.
- Class 4 consists of instances with 1060 or more seconds of social media usage.

For binary predictions, the SMC feature is constructed as follows:

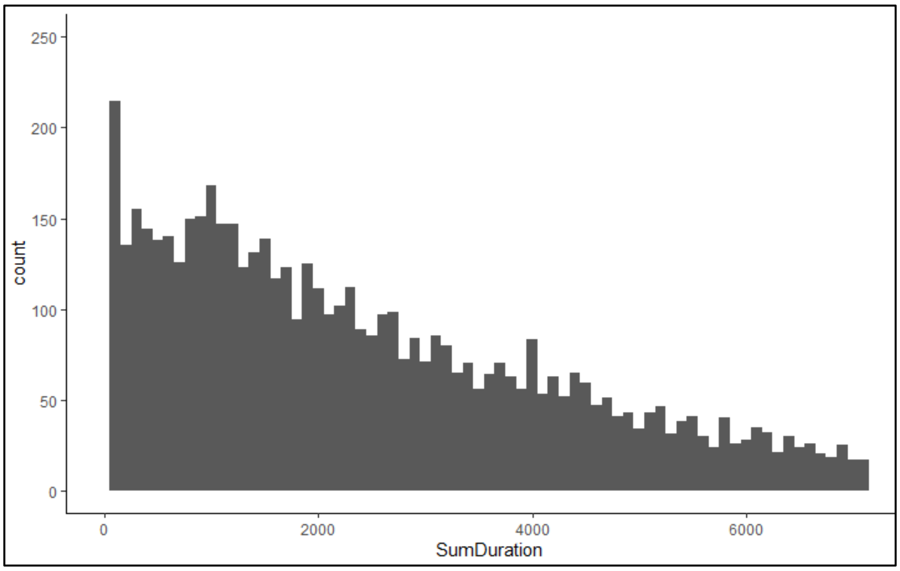
- Class 0 consists of instances with less than 359 seconds of social media usage.
- Class 1 consists of instances with 359 or more seconds of social media usage.

As mentioned in the background, these methods for constructing classes are found in papers by Van Zoonen and Van der Meer (2016) and Singh and Long (2018).

4.4 Data Analysis and Data Transformation

Now that I have the social media classes for every daytime window in the finalized dataset, it is time to reconsider the duration plot. Please find Figure 2 for a plot of the duration feature below:

FIGURE 2: DURATION FEATURE PLOT

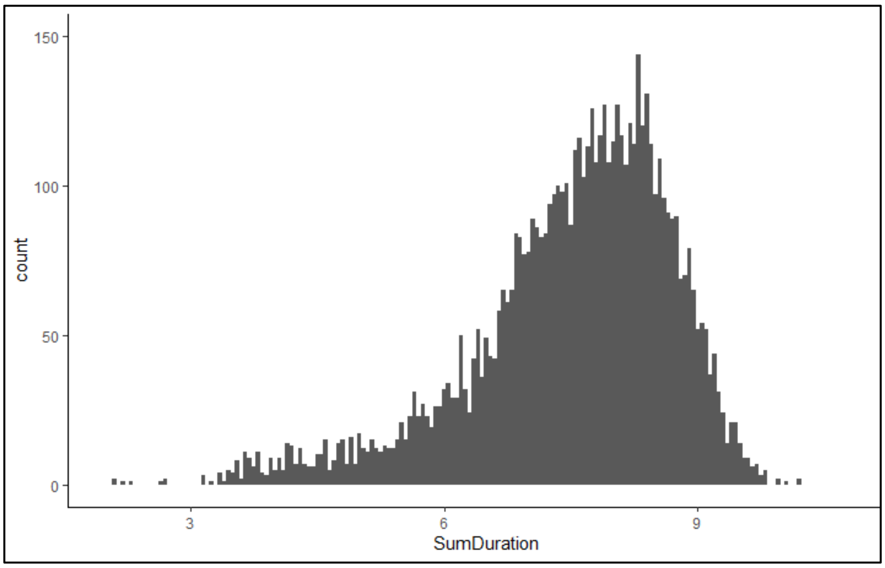


The horizontal axis of Figure 2 shows the total duration in seconds that participants spend on their phones in a certain daytime window. The vertical axis shows how many observations fall within a certain bin. The width of the bins equals 100. Thus, it occurred around 220 times that a participant spent between 0 and 100 seconds on their phone over the timespan of a daytime window.

The tail on the right-hand side is longer than on the left-hand side. For the duration feature, the mean is 3,064 second and the median is 2,255 seconds per daytime window. For the social media feature, the mean is 780.3 seconds and the median is 359.0 seconds per daytime window and the plot is similar to Figure 2. The categorical features show a similar distribution. I conclude that the phone use data is positively skewed.

There are several ways to solve this. Square root transformation, cube root transformation and log transformation are examples of ways to tackle this problem. I have chosen to transform all skewed variables by implementing a “log+1”-transformation. A log transformation is the most appropriate for this specific data, since the data needs a strong transformation given the skewness values for many variables. A log transformation is a relatively strong transformation method and is often useful, since many measurements are lognormally distributed. Since many features are based on the time a participant spends on a certain application category, there are many zero values in the data. To point out an example, there are lots of daytime windows in which a participant does not check an application in the “Auto & Vehicles” category. A regular log transformation does not transform this data as preferred, as the logarithm of zero (*log*(0)) equals minus infinite (-∞). A "log+1"-transformation overcomes this problem by adding one second of duration to all observations before performing the log transformation. This way, the initial zero values become arbitrarily small numbers after transformation. The new distribution is can be found in Figure 3:

FIGURE 3: PLOT OF THE DURATION FEATURE AFTER TRANSFORMATION



As Figure 3 shows, the distribution of the data has improved. The data does not perfectly follow a normal distribution even after transformation, but this is the closest approximation of a normal distribution possible in this case. The data is now ready to be used for feature selection.

4.5 Feature Selection

Correlation feature selection has been utilized before in order to find the most useful features (Nguwi & Loh, 2017). Hall and Smith (1999) initiated this algorithm that still finds applications today. The *cor* function from the *stats* package is used to calculate the correlations between all the variables in the dataset. These correlations are stored in a correlation matrix by using the same function. The *findCorrelation* function from the *caret* package is used to find correlations that are larger than 0.5. A threshold of 0.5 is a generally accepted cut-off point for moderately correlated features. Features that are listed as a result do not necessarily cause problems in the machine learning models. They are simply considered moderately correlated:

```
"gloomy"          "anxious"          "inferior"        "stressed"
"cheerful"      "content"         "SMC_Multiclass"
```

Looking at the listed features, one might consider the names to be intuitive. For example, the *SMC_Multiclass* feature is extremely correlated with *SMC_Binary*, since the classes perfectly correspond to each other. Furthermore, some of the positive and negative mood features are moderately correlated, which is logical. By means of example, when a participant feels stressed, they are also likely to feel more anxious.

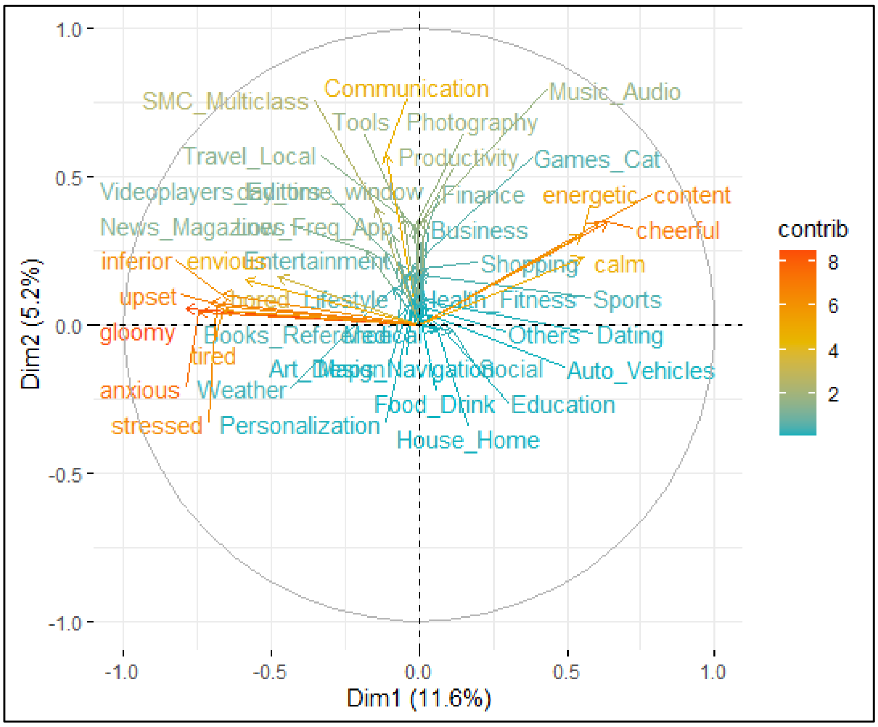
Problems may arise when many of the features turn out to be highly correlated. Here, a cut-off of 0.7 is generally considered acceptable. When checking for features that have a correlation of 0.7 or more, there are no listed features anymore. Therefore, no problems are expected as a result of the correlation analysis.

The *train* function from the *caret* package is used to build a model that uses a simple random forest model to predict the correct SMC using all other variables. The mere goal of implementing this algorithm is to find the variables that most substantially contribute in predicting the right SMC. After training the model, the *varImp* function, also from the *caret* package, is used to assess these most important features. The result is an average relative predictive value per feature in predicting SMC:

```
                     Overall
user_id               787.20
Communication         484.47
Low_Freq_App          395.28
Tools                 217.00
Music_Audio           174.03
Videoplayers_Editors  155.40
cheerful              153.21
Photography           150.37
energetic             146.22
Productivity          141.64
calm                  141.44
content               134.80
day_time_window       133.22
tired                 129.72
stressed              123.33
bored                 114.67
gloomy                 99.53
Travel_Local           87.73
inferior               83.07
anxious                80.77
```

The second way to assess feature importance that I used is principal component analysis (PCA). The *prcomp* function from the *stats* package is used to build a PCA model utilizing all features. The *fviz_eig* and *fviz_pca* functions from the *factoextra* and the *devtools* packages are used to visualize the PCA. These specific functions result in better looking PCA plots than the original *ggbiplot* package. The result can be found in Figure 4.

FIGURE 4: PCA PLOT OF THE DATASET



The mood features have a colour that clearly indicates that these features have a higher contribution to the total variation in the data than the phone use features. Though, the first PC only explains 11.6% of the data and the second PC explains 5.2%. A total of 25 PCs are needed to explain 80% of the variability in the data. A possible explanation is that PCA works better when the features are strongly correlated. If the relationship between the features is weak, PCA does not work well to reduce the data (Van Zoonen & Van der Meer, 2016). The data might not be correlated enough to reduce the dimensionality to only a few features.

After these correlation analyses, the features must be selected accordingly. Three sets of features will be distinguished: a phone feature set, a mood feature set and a set that consists of both kinds of features. Table 1 gives an overview of the feature sets.

TABLE 1: FEATURE SETS

Phone feature set	**Mood feature set**
Low_Freq_App	stressed
Communication	bored
Music_Audio	tired
Productivity	inferior
Tools	gloomy
Travel_Local	energetic
Photography	calm
Videoplayers_Editors	content
user_id	envious
SMC	cheerful
	anxious
	upset
	user_id
	SMC

For the phone feature set, all phone features that are listed as a result of the correlation analysis are used for model training and testing, along with the *user_id* and SMC features. The phone feature set has ten features in it.

In the mood feature set, all twelve mood features are added. The features *envious* and *upset* were not listed as a result of the correlation analysis. Yet, they clearly indicate a colour in the PCA plot that denotes a relatively large contribution to the total variation in the data. Therefore, all mood features are used, along with *user_id* and SMC. As a result, the mood feature set has a total of fourteen features in it.

The combined feature set, that incorporates both kinds of features, consists of all features in the phone and mood set. Since *user_id* and SMC are in both feature sets, there are twenty-two features in the combined feature set. Note that the SMC feature differs per classification problem.

4.6 Machine Learning models

Several machine learning models are implemented to assess their value in predicting SMC from the aggregated dataset. The difference in performance between the feature kinds is interesting from a business perspective. This will be more extensively addressed in the discussion section. For all machine learning models, three sets of features are compared:

1. Only phone features
2. Only mood features
3. Both phone and mood features

Also, there might be differences in performance when a model is used to predict SMC on a binary basis and when a model is used to solve a multi-class problem. Therefore, all machine learning models are evaluation on both these classification problems:

1. Binary classes (split at the median)
2. Four classes (split at the first, second and third quartile)

As a result, all of the models below are evaluated six times. To check the performance of all six variants per model, a test set is used. The proportion of the test set remains equal throughout all variants and model, for consistency reasons. The train proportion of the total amount of data is set to 80% and the test proportion is set to 20%. This split is commonly used in model testing and the split is identical across all models.

This means that the test ensures that evaluation regards the same instances. For this specific dataset, it results in training over 4,908 instances and testing over 1,224 instances. Given the amount of test instances, the test size is large enough to base conclusions on. A detailed overview of the machine learning model performances can be found in the results section.

Naive Bayes

The first machine learning model that is implemented is the Naive Bayes model. The *createDataPartition* function from the *caret* package is used to split the data into a train set and a test set. This method for splitting the train and test sets is only used once, as the other models use the same train and test sets. The cross validation thus occurs randomly between events. After splitting the data, the *naiveBayes* function from the *e1071* package is used to build the model. Consequently, the *predict* function from the *caret* package is used to predict the SMC in the test set by utilizing the trained Naive Bayes model and all features except for the SMC. The *confusionMatrix* function from the *caret* package is used to construct a *confusionMatrix* with accompanying evaluation metrics and descriptive information.

Random Forest

The *randomForest* function from the *randomForest* package is then used to build the random forest model for all sets of features and both classification problems. In RStudio, the standard number of trees is 500 and the variables tried at each split for this data is fifteen.

After the model is built, the predict function is used again to predict the SMC feature in the test set. Again, the *confusionMatrix* function from the *caret* package is used to build a confusion matrix with descriptive information and metrics.

Support Vector Machine

The last machine learning model that is used is the Support Vector Machine (SVM). To build the models, the *svm* function from the *e1071* package is used. The train and test sets are both scaled using the *scale* function that is standard in R. This results in datasets that are more evenly distributed and more appropriate for building the SVM model. The chosen type of SVM is C-classification. The “C” is a regularization parameter. It controls the trade-off between errors in the training and test set. It tries to generalize the model to the test data instead of only the training data. Furthermore, the kernel is set to radial. This enables the model to perform a “kernel trick”. The radial basis function kernel in SVMs overcomes the problem of linearity by avoiding the mapping that is needed to learn a non-linear function or decision boundary. This way, the decision boundary can be non-linear and generally performs better in classification tasks. After building the model, the *predict* function is used once again to predict the SMC feature in the test set. Also, a confusion matrix is made using the *confusionMatrix* function.

5. Results

In this section, the performances of the classification models will be assessed. The aim of the machine learning models was to correctly predict in which social media class (SMC) a certain participant belongs and to predict the positive class as correctly as possible. This is done on a daytime window basis, where days are split up in four daytime windows. These daytime window are explained in the beginning of the experimental setup section. The evaluation metrics that are used are accuracy and recall/sensitivity. These metrics are assessed six times per machine learning algorithm. All models are trained on phone use features and mood features, separately, as well as both simultaneously. Furthermore, the models are trained to predict the correct SMC on a binary as well as on a multiclass basis. The classes are based on quartiles that contain an equal number of instances. As a result, the classes are always balanced. Therefore, the benchmark in this study is always 50% for the binary models and 25% for the multiclass models that consist of four classes. This goes for both the accuracy and the recall metrics. This section gives an overview of the performances of the models. Please see Appendix A for a more detailed commentary of the results.

The following two tables give a summary of the total performance of the models. In the columns, the ‘P’ stands for phone usage features, the ‘M’ stands for mood features and ‘P + M’ stands for both kind of features. In the rows, ‘NB’ stands for Naive Bayes, ‘RF’ stands for random forest and ‘SVM’ stands for support vector machine.

TABLE 2: ACCURACY PERFORMANCE OF ALL MODELS

Accuracy	Binary model			Multiclass model		
	P	M	P + M	P	M	P + M
NB	0.5771	0.5616	0.6188	0.3252	0.3015	0.3301
RF	0.7270	0.7172	0.7406	0.3866	0.3705	0.4040
SVM	0.6416	0.6073	0.6588	0.3881	0.3578	0.4404

Accuracy

The first metric to evaluate is accuracy. Remarkable for the accuracy metric is that for every different machine learning algorithm, the order of performance is the same: a combined model that includes both phone and mood features performs best, then comes a model that only incorporates phone features and a mood feature model always performs worst. This seems intuitive, since it is assumable that phone features might have more explanatory value in predicting the correct SMC than mood features. By means of example, the time that participants spend on communication applications might have more predictive value than their level of stress.

All models outperform their benchmarks, yet there are some differences. For the binary models, the RF algorithm is the best performing algorithm when it comes to accuracy, followed by the SVM. Regarding the multiclass models, the RF and SVM models are comparable. The NB classifier performs worst in all cases.

Mood features do have additional value in predicting the correct SMC, yet only when combined with phone features. This additional value varies between two percent to around five percent at the most.

TABLE 3: RECALL PERFORMANCE OF ALL MODELS

Recall	Binary model			Multiclass model		
	P	M	P + M	P	M	P + M
NB	0.3475	0.4388	0.4584	0.2320	0.2549	0.3039
RF	0.6975	0.7051	0.7410	0.1065	0.0960	0.1043
SVM	0.6232	0.6199	0.6378	0.4542	0.3660	0.4804

Recall

The recall metric denotes which proportion of positive classes is also classified as such. For the binary models, the positive class is ‘1’ and for multiclass models, the positive class is ‘4’. These positive classes correspond to the highest degree of social media usage of all classes. As mentioned, this metric is interesting from a societal perspective, since predicting problematic social media usage is an important phase in preventing it.

Other than for the accuracy metric, the order of the performances of the models in terms of features is not always the same. Although a combined model that includes both phone and mood features always performs best in terms of recall, models that incorporate either feature kind differ in their spot in the ranking.

The NB classifier fails to outperform the 50% benchmark for all binary models, it just does so for two out of three multiclass models. The RF and SVM algorithms both outperform the benchmark when it comes to binary models. Only the SVM consistently performs better than the benchmark for multiclass models. The RF algorithm performs very poorly in terms of recall when using multiclass model variants.

Error analysis

An interesting note on the error distribution of the machine learning models can be made. In terms of errors, the confusion matrices in Appendix A show a trend. A Type 1 error occurs when the models predict a positive value, yet the actual value turns out to be negative. A Type 2 error occurs when the model predict a negative value, yet the actual value turns out to be positive. The Type 2 error occurs more often than the Type 1 error. For the binary models, the models often predict the base class when the actual value turns out to be positive (1). For the multiclass models, the models very often predict either 1, 2 or 3 when the positive value is actually correct, which denotes the class 4. Please see Appendix A for the tables on which these claims are based.

6. Discussion

General performance and the value of feature kinds

The goal of this study was to predict the correct social media class (SMC) of participants by utilizing machine learning methods. Both binary and multiclass models were built and evaluated on accuracy and recall. Also, a distinction was made between phone and mood features. Literature shows that mood survey data is more expensive than phone tracking data. Therefore, a trade-off between these data types is made. The feature selection stage showed that features from both the phone tracking data as well as the mood survey data are important when predicting the correct SMC.

Three machine learning algorithms were assessed on their predictive performance: Naive Bayes, random forest and support vector machine. Overall, the random forest and support vector machine algorithms perform decently. The Naive Bayes classifier clearly performed worst. All models outperformed their benchmark in terms of accuracy. Also, an obvious trend in the results was the order of performances for the models. Regarding accuracy, a model that used both kind of features always performed best, followed by a model that only used phone features. A mood feature model always performed worst. An important remark is that the combined model, which uses both kind of features, only outperforms a more simple phone feature model by a couple percentage points. Since phone tracking data is cheaper than mood survey data, the additional value of mood data does not outweigh its costs.

Both these findings do not occur for the recall metric. As mentioned, the recall metric is very useful when it comes to predicting problematic social media usage. When a model accurately predicts the problematic highest quartile of social media usage correctly, this is an important phase in preventing problematic social media usage. Several possibilities to prevent social media usage are pointed out in the background section. Unfortunately, the models do not provide a sufficiently large recall to effectively predict problematic instances. For the multiclass models, the recall does not exceed fifty percent. Altogether, from the results it follows that SMCs are difficult to predict when using a combination of phone and mood features.

Other influential factors

The literature gives several explanations for these results. Wood, Bukowski and Lis (2016) suggested that social media might have more positive outcomes for males than for females. Males might be building social skills, whereas females experience more negative ramifications that result in lower self-esteem. This statement is underwritten by Orben, Dienlin and Przybylski (2019). Since the sample in this study consists of non-identifiable natural persons, it is possible that a majority of the sample is male.

Furthermore, it has been pointed out that ease of use, usefulness and satisfaction are important factors when it comes to continuance of social media usage (Idemudia, Raisinghani & Samual-Ojo, 2016). The authors show that some social media platforms might no longer satisfy the aforementioned needs that are required for the platforms to be continued to use. More specifically, there might be certain applications that are considered social media in the current study, that to not reach the point of problematic use. By means of example, Facebook usage has evolved in such a way that it is less problematic nowadays. This means that individuals compare themselves less to their social environment than before, since Instagram has taken over a large proportion of that aspect from Facebook. In future research, the current trends in social media platforms should be assessed. Knowing which social media platforms are used to a problematic extent is key.

Apart from that, an important relationship regarding this topic is addressed. Kuang-Tsan and Fu-Yuan (2017) showed that academic stress has an influence on mobile phone addiction for Taiwanese students. The relationship between negative mood characteristics and smart phone use for students has been addressed by other researchers as well (Lee et al, 2017). It is useful to study whether this relationship can be generalized to other age categories. It is assumable that academic stress is an important factor in the relationship between stress and social media usage, which is why individuals in other age categories do not experience it.

Lastly, there might simply be no connection between social media usage and the utilized features. Orben, Dienlin and Przybylski (2019) claimed that social media usage is not a strong predictor for life satisfaction. El-Badawy and Hashem (2015) make a similar statement, indicating that social media usage does not predict academic performance. Another paper shows that passive smartphone data is not suitable for predicting mood (Pratap et al., 2018). The current study tries to predict SMC from mood and smartphone data, which is definitely not the same, since the dependent variable and the predictors are switched. Still, it might be a valid argument that the relationships are overestimated, since multiple studies claim that the connections do not exist.

Contribution to literature

This study contributes to existing literature, since no research has aimed to predict social media classes by using a machine learning approach. Also, a trade-off between phone and mood features has not been made before. The results do not imply that problematic social media use can be predicted, as the recall metric does not show good results. This would mean that, for now, businesses cannot exploit the results. Though, there are opportunities for future research.

7. Conclusion

Predicting the correct SMC was the key topic of this study. The pre-processing phase was a large part of the total analysis, where data transformation was a necessary step. Mood features are significantly more expensive than phone features. These mood features, when combined with phone features, improve every model in terms of both accuracy and recall, yet only to a small extent. Therefore, the additional predictive value of incorporating mood features is unlikely to outweigh the costs. Furthermore, recall was an important metric to assess the potential prevention of problematic social media usage. Unfortunately, the machine learning models did not perform well enough to make a sound claim about problematic social media prediction. The impact on businesses is not as large as preferable, yet it might be when more research is conducted.

Future research could focus on quantifying the difference between males and females when it comes to the negative consequences of social media usage. When predicting social media classes, this difference might partially explain the results. Also, the current trends on social media platforms should always be taken into account. Some social media platforms might not be used to a problematic extent anymore. Generalizing claims on the relationship between social media and mood to other age categories can also be of interest. Lastly, there might even be a possibility that the mentioned connections between variables do not exist to an exploitable extent, or do not exist at all.

References

Agarwal, N. K., & Lu, W. (2019). Toward Effective Use and Balance: Investigating Pros and Cons of Smartphone Use in an Interview Study. *ASIS&T* , *82*, 594–596. https://doi.org/10.1002/pra2.00100

Bakker, D., & Rickard, N. (2019). Engagement with a cognitive behavioural therapy mobile phone app predicts changes in mental health and wellbeing: MoodMission. *Australian Psychologist*, *54*(4), 245–260. https://doi.org/10.1111/ap.12383

Berry, N., Emsley, R., Lobban, F., & Bucci, S. (2018). Social media and its relationship with mood, self-esteem and paranoia in psychosis. *Acta Psychiatrica Scandinavica*, *138*(6), 558–570. https://doi.org/10.1111/acps.12953

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

El-Badawy, T. A., & Hashem, Y. (2014). The Impact of Social Media on the Academic Development of School Students. *International Journal of Business Administration*, *6*(1), 1. https://doi.org/10.5430/ijba.v6n1p46

Evgeniou, T., & Pontil, M. (2001). Support Vector Machines: Theory and Applications. *Machine Learning and Its Applications*, 249–257. https://doi.org/10.1007/3-540-44673-7_12

Griffioen, N., Van Rooij, M. M. J. W., Lichtwarck-Aschoff, A., & Granic, I. (2020). A Stimulated Recall Method for the Improved Assessment of Quantity and Quality of Social Media Use. *Journal of Medical Internet Research*, *22*(1), e15529. https://doi.org/10.2196/15529

Gupta, N., & Irwin, J. D. (2016). In-class distractions: The role of Facebook and the primary learning task. *Computers in Human Behavior*, *55*, 1165–1178. https://doi.org/10.1016/j.chb.2014.10.022

Hand, D. J., & Yu, K. (2001). Idiot's Bayes: Not So Stupid after All? *International Statistical Review / Revue Internationale de Statistique*, *69*(3), 385. https://doi.org/10.2307/1403452

Home. (2009). Retrieved June 15, 2020, from https://ethicadata.com/

Kuang-Tsan, C., & Fu-Yuan, H. (2017). Study on Relationship Among University Students' Life Stress, Smart Mobile Phone Addiction, and Life Satisfaction. *Journal of Adult Development*, *24*(2), 109–118. https://doi.org/10.1007/s10804-016-9250-9

Lau, W. W. F. (2017). Effects of social media usage and social media multitasking on the academic performance of university students. *Computers in Human Behavior*, *68*, 286–291. https://doi.org/10.1016/j.chb.2016.11.043

Lee, S., Kim, M. W., McDonough, I. M., Mendoza, J. S., & Kim, M. S. (2017). The Effects of Cell Phone Use and Emotion-regulation Style on College Students' Learning. *Applied Cognitive Psychology*, *31*(3), 360–366. https://doi.org/10.1002/acp.3323

Lin, L. yi, Sidani, J. E., Shensa, A., Radovic, A., Miller, E., Colditz, J. B., … Primack, B. A. (2016). Association between Social Media Use and Depression among U.S. Young Adults. *Depression and Anxiety*, *33*(4), 323–331. https://doi.org/10.1002/da.22466

Luhmann, M. (2017). Using Big Data to study subjective well-being. *Current Opinion in Behavioral Sciences*, *18*, 28–33. https://doi.org/10.1016/j.cobeha.2017.07.006

Nguwi, Y.-Y., & Loh, P. R. (2017). Comparative Study of Feature Selection and Classification for Problematic Mobile Phone Use (PMPU). *Journal of Technology in Behavioral Science*, *2*(3–4), 121–128. https://doi.org/10.1007/s41347-017-0023-5

Orben, A., Dienlin, T., & Przybylski, A. K. (2019). Social media's enduring effect on adolescent life satisfaction. *Proceedings of the National Academy of Sciences*, *116*(21), 10226–10228. https://doi.org/10.1073/pnas.1902058116

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The*

London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Pratap, A., Atkins, D. C., Renn, B. N., Tanana, M. J., Mooney, S. D., Anguera, J. A., & Areán, P. A. (2018). The accuracy of passive phone sensors in predicting daily mood. *Depression and Anxiety*, *36*(1), 72–81. https://doi.org/10.1002/da.22822

Robinson, A., Bonnette, A., Howard, K., Ceballos, N., Dailey, S., Lu, Y., & Grimes, T. (2019). Social comparisons, social media addiction, and social interaction: An examination of specific social media behaviors related to major depressive disorder in a millennial population. *Journal of Applied Biobehavioral Research*, *24*(1), 1–14. https://doi.org/10.1111/jabr.12158

Satici, S. A. (2018). Facebook Addiction and Subjective Well-Being: a Study of the Mediating Role of Shyness and Loneliness. *International Journal of Mental Health and Addiction*, *17*(1), 41–55. https://doi.org/10.1007/s11469-017-9862-8

Servia-Rodríguez, S., Rachuri, K. K., Mascolo, C., Rentfrow, P. J., Lathia, N., & Sandstrom, G. M. (2017). Mobile Sensing at the Service of Mental Well-being. *Proceedings of the 26th International Conference on World Wide Web - WWW '17*. https://doi.org/10.1145/3038912.3052618

Singh, V. K., & Long, T. (2018). Automatic assessment of mental health using phone metadata. *Proceedings of the Association for Information Science and Technology*, *55*(1), 450–459. https://doi.org/10.1002/pra2.2018.14505501049

Treem, J. W., Dailey, S. L., Pierce, C. S., & Biffl, D. (2016). What We Are Talking About When We Talk About Social Media: A Framework for Study. *Sociology Compass*, *10*(9), 768–784. https://doi.org/10.1111/soc4.12404

Van Zoonen, W., & Van der Meer, T., G. L. A. (2016). Social media research: The application of supervised machine learning in organizational communication research. *Computers in Human Behavior*, *63*, 132–141. https://doi.org/10.1016/j.chb.2016.05.028

Whiting, A., & Williams, D. (2013). Why people use social media: a uses and gratifications approach. *Qualitative Market Research: An International Journal*, *16*(4), 362–369. https://doi.org/10.1108/qmr-06-2013-0041

Wood, M. A., Bukowski, W. M., & Lis, E. (2015). The Digital Self: How Social Media Serves as a Setting that Shapes Youth's Emotional Experiences. *Adolescent Research Review*, *1*(2), 163–173. https://doi.org/10.1007/s40894-015-0014-8

Zha, Z.-J., Mei, T., & El Saddik, A. (2016). Social media analytics and learning. *Neurocomputing*, *172*, 1–2. https://doi.org/10.1016/j.neucom.2015.04.082

Appendix A: Detailed Results

Naive Bayes

The three binary models performed as follows:

FIGURE 5: BINARY NB ONLY PHONE FEATURES

```
preds   0   1
    0 494 400
    1 118 213

               Accuracy : 0.5771
                 95% CI : (0.5489, 0.605)
    No Information Rate : 0.5004
    P-Value [Acc > NIR] : 4.321e-08

                  Kappa : 0.1546

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.3475
            Specificity : 0.8072
         Pos Pred Value : 0.6435
         Neg Pred Value : 0.5526
             Prevalence : 0.5004
         Detection Rate : 0.1739
   Detection Prevalence : 0.2702
      Balanced Accuracy : 0.5773

       'Positive' Class : 1
```

FIGURE 6: BINARY NB ONLY MOOD FEATURES

```
preds   0   1
    0 419 344
    1 193 269

               Accuracy : 0.5616
                 95% CI : (0.5333, 0.5897)
    No Information Rate : 0.5004
    P-Value [Acc > NIR] : 1.012e-05

                  Kappa : 0.1234

 Mcnemar's Test P-Value : 9.609e-11

            Sensitivity : 0.4388
            Specificity : 0.6846
         Pos Pred Value : 0.5823
         Neg Pred Value : 0.5491
             Prevalence : 0.5004
         Detection Rate : 0.2196
   Detection Prevalence : 0.3771
      Balanced Accuracy : 0.5617

       'Positive' Class : 1
```

FIGURE 7: BINARY NB BOTH KIND OF FEATURES

```
preds   0   1
    0 477 332
    1 135 281

               Accuracy : 0.6188
                 95% CI : (0.5909, 0.6461)
    No Information Rate : 0.5004
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2378

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4584
            Specificity : 0.7794
         Pos Pred Value : 0.6755
         Neg Pred Value : 0.5896
             Prevalence : 0.5004
         Detection Rate : 0.2294
   Detection Prevalence : 0.3396
      Balanced Accuracy : 0.6189

       'Positive' Class : 1
```

What stands out for these three figures is that the base, the zero class, is guessed too often. At least 800 out of 1,225 examples are predicted to be 0, where the actual number is 612. Furthermore, the phone features in Figure 5 provide an accuracy of 57.71% and a sensitivity of 34.75%. The results differ when only mood features are used, the accuracy then becomes 56.16% and the sensitivity 43.88%, as can be found in Figure 6. However, Figure 7 shows that combining both kind of features results in both a higher accuracy and sensitivity; 61.88% and 45.84%. Phone features in a binary Naive Bayes classifier provide a slightly higher accuracy than mood features, whereas mood features result in a higher sensitivity. All models perform better than the benchmark when it comes to accuracy. The sensitivity is lower than the benchmark.

The three multiclass models performed as follows:

FIGURE 8: MULTICLASS NB ONLY PHONE FEATURES

```
preds   1   2   3   4
    1 267 222 206 203
    2  11  26  23  15
    3   4  23  34  17
    4  23  36  43  71

Overall Statistics

               Accuracy : 0.3252
                 95% CI : (0.299, 0.3522)
    No Information Rate : 0.2508
    P-Value [Acc > NIR] : 3.241e-09

                  Kappa : 0.1009

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.8754  0.08469  0.11111  0.23203
Specificity            0.3134  0.94656  0.95207  0.88889
Pos Pred Value         0.2973  0.34667  0.43590  0.41040
Neg Pred Value         0.8834  0.75544  0.76265  0.77640
Prevalence             0.2492  0.25082  0.25000  0.25000
Detection Rate         0.2181  0.02124  0.02778  0.05801
Detection Prevalence   0.7337  0.06127  0.06373  0.14134
Balanced Accuracy      0.5944  0.51563  0.53159  0.56046
```

FIGURE 9: MULTICLASS NB ONLY MOOD FEATURES

```
preds   1   2   3   4
    1 151 128  97  99
    2  57  69  71  59
    3  46  53  71  70
    4  51  57  67  78

Overall Statistics

               Accuracy : 0.3015
                 95% CI : (0.2759, 0.328)
    No Information Rate : 0.2508
    P-Value [Acc > NIR] : 3.462e-05

                  Kappa : 0.0688

 Mcnemar's Test P-Value : 8.686e-12

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.4951  0.22476  0.23203  0.25490
Specificity            0.6474  0.79607  0.81590  0.80937
Pos Pred Value         0.3179  0.26953  0.29583  0.30830
Neg Pred Value         0.7944  0.75413  0.76118  0.76519
Prevalence             0.2492  0.25082  0.25000  0.25000
Detection Rate         0.1234  0.05637  0.05801  0.06373
Detection Prevalence   0.3881  0.20915  0.19608  0.20670
Balanced Accuracy      0.5713  0.51041  0.52397  0.53214
```

FIGURE 10: MULTICLASS NB WITH BOTH FEATURE KINDS

```
preds   1   2   3   4
    1 231 194 193 161
    2  12  27  17  14
    3  23  50  53  38
    4  39  36  43  93

Overall Statistics

               Accuracy : 0.3301
                 95% CI : (0.3037, 0.3572)
    No Information Rate : 0.2508
    P-Value [Acc > NIR] : 3.245e-10

                  Kappa : 0.1073

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.7574  0.08795   0.1732  0.30392
Specificity            0.4037  0.95311   0.8791  0.87146
Pos Pred Value         0.2965  0.38571   0.3232  0.44076
Neg Pred Value         0.8337  0.75737   0.7613  0.78973
Prevalence             0.2492  0.25082   0.2500  0.25000
Detection Rate         0.1887  0.02206   0.0433  0.07598
Detection Prevalence   0.6364  0.05719   0.1340  0.17239
Balanced Accuracy      0.5805  0.52053   0.5261  0.58769
```

The phone features, again, seem to better predict SMC than mood features. The combination of both gives a slightly better result than either of the kind of features. The accuracy is higher than the benchmark for all three model variants. In terms of sensitivity, the phone model underperforms the benchmark and the mood feature and combined models outperform the benchmark.

Random Forest

The three binary model variants performed as follows:

FIGURE 11: BINARY RF PHONE FEATURES

```
pred   1   2
   1 641 224
   2 278 696

               Accuracy : 0.727
                 95% CI : (0.706, 0.7473)
    No Information Rate : 0.5003
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.454

 Mcnemar's Test P-Value : 0.01801

            Sensitivity : 0.6975
            Specificity : 0.7565
         Pos Pred Value : 0.7410
         Neg Pred Value : 0.7146
             Prevalence : 0.4997
         Detection Rate : 0.3486
   Detection Prevalence : 0.4704
      Balanced Accuracy : 0.7270

       'Positive' Class : 1
```

FIGURE 12: BINARY RF WITH MOOD FEATURES

```
pred   1   2
   1 648 249
   2 271 671

               Accuracy : 0.7172
                 95% CI : (0.696, 0.7377)
    No Information Rate : 0.5003
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.4345

 Mcnemar's Test P-Value : 0.3571

            Sensitivity : 0.7051
            Specificity : 0.7293
         Pos Pred Value : 0.7224
         Neg Pred Value : 0.7123
             Prevalence : 0.4997
         Detection Rate : 0.3524
   Detection Prevalence : 0.4878
      Balanced Accuracy : 0.7172

       'Positive' Class : 1
```

FIGURE 13: BINARY RF BOTH FEATURE KINDS

```
pred   1   2
   1 681 239
   2 238 681

               Accuracy : 0.7406
                 95% CI : (0.7199, 0.7605)
    No Information Rate : 0.5003
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.4812

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.7410
            Specificity : 0.7402
         Pos Pred Value : 0.7402
         Neg Pred Value : 0.7410
             Prevalence : 0.4997
         Detection Rate : 0.3703
   Detection Prevalence : 0.5003
      Balanced Accuracy : 0.7406

       'Positive' Class : 1
```

The binary random forest model performs much better than the binary Naive Bayes model, both in terms of accuracy and sensitivity. Again, the phone feature model performs a bit better than the mood feature model in terms of accuracy, yet slightly worse in terms of sensitivity. The overall accuracy is around 72% and the overall sensitivity is around 71%.

The random forest model algorithm not seem to encounter a similar problem as the Naive Bayes algorithm, predicting the '0' class in an obvious majority of the cases.

The three multiclass models performed as follows:

FIGURE 14: MULTICLASS RF PHONE FEATURES

```
pred   1   2   3   4
   1 112   8   3   0
   2 281 240 139  71
   3  65 211 310 340
   4   0   2   8  49

Overall Statistics

               Accuracy : 0.3866
                 95% CI : (0.3643, 0.4093)
    No Information Rate : 0.2507
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.1819

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity           0.24454   0.5206   0.6739  0.10652
Specificity           0.99203   0.6437   0.5533  0.99275
Pos Pred Value        0.91057   0.3283   0.3348  0.83051
Neg Pred Value        0.79837   0.8005   0.8357  0.76910
Prevalence            0.24905   0.2507   0.2501  0.25014
Detection Rate        0.06090   0.1305   0.1686  0.02664
Detection Prevalence  0.06688   0.3975   0.5035  0.03208
Balanced Accuracy     0.61829   0.5821   0.6136  0.54964
```

FIGURE 15: MULTICLASS RF MOOD FEATURES

```
pred   1   2   3   4
   1 105   4   0   0
   2 273 234 145  71
   3  80 218 298 344
   4   0   5  17  44

Overall Statistics

               Accuracy : 0.3705
                 95% CI : (0.3484, 0.3931)
    No Information Rate : 0.2508
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.1602

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity           0.22926   0.5076   0.6478  0.09586
Specificity           0.99710   0.6449   0.5341  0.98405
Pos Pred Value        0.96330   0.3237   0.3170  0.66667
Neg Pred Value        0.79584   0.7964   0.8196  0.76580
Prevalence            0.24918   0.2508   0.2503  0.24973
Detection Rate        0.05713   0.1273   0.1621  0.02394
Detection Prevalence  0.05930   0.3934   0.5114  0.03591
Balanced Accuracy     0.61318   0.5762   0.5910  0.53995
```

FIGURE 16: MULTICLASS RF BOTH FEATURE KINDS

```
pred   1   2   3   4
   1 113   2   2   0
   2 295 265 134  60
   3  50 192 317 352
   4   0   2   7  48

Overall Statistics

               Accuracy : 0.404
                 95% CI : (0.3815, 0.4269)
    No Information Rate : 0.2507
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2051

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity           0.24672   0.5748   0.6891   0.1043
Specificity           0.99710   0.6451   0.5693   0.9935
Pos Pred Value        0.96581   0.3515   0.3480   0.8421
Neg Pred Value        0.79965   0.8194   0.8459   0.7688
Prevalence            0.24905   0.2507   0.2501   0.2501
Detection Rate        0.06145   0.1441   0.1724   0.0261
Detection Prevalence  0.06362   0.4100   0.4954   0.0310
Balanced Accuracy     0.62191   0.6100   0.6292   0.5489
```

What is remarkable about these results is that the sensitivity of all model variants is extremely low. Predicting the problematic SMC '4' only succeeds in around 10% of the cases in which a participant actually belongs to this class. This sensitivity percentage is around 15% lower than the benchmark of 25%. The accuracy is on average 38% for these multiclass models. This is 13% above the benchmark. A model with only phone features performs a bit better than a model with only mood features. The combination performs slightly better.

Support Vector Machine

The three binary models performed as follows:

FIGURE 17: BINARY SVM PHONE FEATURES

```
y_pred   0   1
     0 404 231
     1 208 382

               Accuracy : 0.6416
                 95% CI : (0.6141, 0.6685)
    No Information Rate : 0.5004
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.2833

 Mcnemar's Test P-Value : 0.2937

            Sensitivity : 0.6232
            Specificity : 0.6601
         Pos Pred Value : 0.6475
         Neg Pred Value : 0.6362
             Prevalence : 0.5004
         Detection Rate : 0.3118
   Detection Prevalence : 0.4816
      Balanced Accuracy : 0.6416

       'Positive' Class : 1
```

FIGURE 18: BINARY SVM MOOD FEATURES

```
y_pred   0   1
     0 364 233
     1 248 380

               Accuracy : 0.6073
                 95% CI : (0.5794, 0.6348)
    No Information Rate : 0.5004
    P-Value [Acc > NIR] : 3.557e-14

                  Kappa : 0.2147

 Mcnemar's Test P-Value : 0.5232

            Sensitivity : 0.6199
            Specificity : 0.5948
         Pos Pred Value : 0.6051
         Neg Pred Value : 0.6097
             Prevalence : 0.5004
         Detection Rate : 0.3102
   Detection Prevalence : 0.5127
      Balanced Accuracy : 0.6073

       'Positive' Class : 1
```

FIGURE 19: BINARY SVM BOTH FEATURE KINDS

```
y_pred   0   1
     0 416 222
     1 196 391

               Accuracy : 0.6588
                 95% CI : (0.6315, 0.6853)
    No Information Rate : 0.5004
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.3176

 Mcnemar's Test P-Value : 0.2214

            Sensitivity : 0.6378
            Specificity : 0.6797
         Pos Pred Value : 0.6661
         Neg Pred Value : 0.6520
             Prevalence : 0.5004
         Detection Rate : 0.3192
   Detection Prevalence : 0.4792
      Balanced Accuracy : 0.6588

       'Positive' Class : 1
```

The binary SVM slightly underperforms the random forest in terms of both accuracy and sensitivity, but still outperforms the benchmark. The phone feature model outperforms the mood model in both accuracy and sensitivity. The binary Naive Bayes and random forest algorithms showed that the mood feature model outperformed the phone feature model in terms of sensitivity. Yet, these differences are so minor that no real conclusions can be based on them. Might the model be trained again with the same data, disabling the *set.seed* function that is standard in R, then the results might differ slightly in favor of another model variant.

The three multiclass model performed as follows:

FIGURE 20: MULTICLASS SVM PHONE FEATURES

```
y_pred   1   2   3   4
     1 170  98  82  65
     2  62  99  74  51
     3  30  64  67  51
     4  43  46  83 139

Overall Statistics

               Accuracy : 0.3881
                 95% CI : (0.3607, 0.416)
    No Information Rate : 0.2508
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.1842

 Mcnemar's Test P-Value : 3.99e-08

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.5574  0.32248  0.21895   0.4542
Specificity            0.7334  0.79607  0.84205   0.8126
Pos Pred Value         0.4096  0.34615  0.31604   0.4469
Neg Pred Value         0.8331  0.77825  0.76383   0.8171
Prevalence             0.2492  0.25082  0.25000   0.2500
Detection Rate         0.1389  0.08088  0.05474   0.1136
Detection Prevalence   0.3391  0.23366  0.17320   0.2541
Balanced Accuracy      0.6454  0.55927  0.53050   0.6334
```

FIGURE 21: MULTICLASS SVM MOOD FEATURES

```
y_pred   1   2   3   4
     1 153  80  70  54
     2  81 110 101  83
     3  29  72  63  57
     4  42  45  72 112

Overall Statistics

               Accuracy : 0.3578
                 95% CI : (0.3309, 0.3854)
    No Information Rate : 0.2508
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.1438

 Mcnemar's Test P-Value : 2.333e-06

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.5016  0.35831  0.20588   0.3660
Specificity            0.7780  0.71101  0.82789   0.8268
Pos Pred Value         0.4286  0.29333  0.28507   0.4133
Neg Pred Value         0.8247  0.76796  0.75773   0.7964
Prevalence             0.2492  0.25082  0.25000   0.2500
Detection Rate         0.1250  0.08987  0.05147   0.0915
Detection Prevalence   0.2917  0.30637  0.18056   0.2214
Balanced Accuracy      0.6398  0.53466  0.51688   0.5964
```

FIGURE 22: MULTICLASS SVM BOTH FEATURE KINDS

```
y_pred   1   2   3   4
     1 172  69  64  38
     2  66 133  86  55
     3  32  67  87  66
     4  35  38  69 147

Overall Statistics

               Accuracy : 0.4404
                 95% CI : (0.4123, 0.4687)
    No Information Rate : 0.2508
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.2538

 Mcnemar's Test P-Value : 0.01181

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.5639   0.4332  0.28431   0.4804
Specificity            0.8139   0.7743  0.82026   0.8453
Pos Pred Value         0.5015   0.3912  0.34524   0.5087
Neg Pred Value         0.8490   0.8032  0.77469   0.8299
Prevalence             0.2492   0.2508  0.25000   0.2500
Detection Rate         0.1405   0.1087  0.07108   0.1201
Detection Prevalence   0.2802   0.2778  0.20588   0.2361
Balanced Accuracy      0.6889   0.6037  0.55229   0.6629
```

The multiclass SVM outperforms the multiclass Naive Bayes model in both terms of accuracy and recall. Compared to the multiclass random forest, the SVM outperforms in terms of sensitivity and performs comparable in terms of accuracy. The average accuracy is around 40%, where the order of well performing models follows the same trend as for the Naive Bayes and random forest models. The average recall is around 43%, where the same order exists.