Computers in Human Behavior 126 (2022) 106990

Contents lists available at ScienceDirect

Computers in Human Behavior

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

Age and gender in language, emoji, and emoticon usage in instant messages

Timo K. Koch [a], Peter Romero [b], Clemens Stachl [c,*]

[a] Department of Psychology, Psychological Methods and Assessment, Ludwig-Maximilians-Universität München, Leopoldstr. 13, 80802, Munich, Germany
[b] Graduate School of Economics, Advanced Cognition Labs, Keio University, 2-15-45 Mita, Minato-ku, Tokyo, 108-8345, Japan
[c] Institute of Behavioral Science and Technology, University of St. Gallen, Torstrasse 25, CH-9000, St. Gallen, Switzerland

ARTICLE INFO

Keywords:
Age
Gender
Author profiling
Instant messages
Machine learning
Digital footprints

ABSTRACT

Text is one of the most prevalent types of digital data that people create as they go about their lives. Digital footprints of people's language usage in social media posts were found to allow for inferences of their age and gender. However, the even more prevalent and potentially more sensitive text from instant messaging services has remained largely uninvestigated. We analyze language variations in instant messages with regard to individual differences in age and gender by replicating and extending the methods used in prior research on social media posts. Using a dataset of 309,229 WhatsApp messages from 226 volunteers, we identify unique age- and gender-linked language variations. We use cross-validated machine learning algorithms to predict volunteers' age ($MAE_{Md} = 3.95$, $r_{Md} = 0.81$, $R^2_{Md} = 0.49$) and gender ($Accuracy_{Md} = 85.7\%$, $F1_{Md} = 0.67$, $AUC_{Md} = .82$) significantly above baseline-levels and identify the most predictive language features. We discuss implications for psycholinguistic theory, present opportunities for application in author profiling, and suggest methodological approaches for making predictions from small text data sets. Given the recent trend towards the dominant use of private messaging and increasingly weaker user data protection, we highlight rising threats to individual privacy rights in instant messaging.

1. Introduction

When texting a friend on WhatsApp, posting on Facebook, tweeting on Twitter, or writing a blog post, we inevitably leave behind digital footprints in the form of text data. Research in the domain of *author profiling* has shown that language characteristics of Facebook status updates (Jaidka et al., 2018; Sap et al., 2014; Schwartz et al., 2013), tweets (Bamman et al., 2014; Burger et al., 2011; Jaidka et al., 2018; Rao et al., 2010; Sap et al., 2014), and blog posts (Argamon et al., 2007; Sap et al., 2014; Schler et al., 2006) allow for the accurate inference of the authors' age and gender. Moreover, these social media studies extended the theory of gender- and age-linked language variations (Park et al., 2016). Instant messaging services (e.g., WhatsApp, Facebook Messenger, WeChat) also produce vast amounts of digital footprints every day, but have rarely been investigated in language studies. Unlike data from social media platforms, like Facebook, Twitter, and Reddit, text from instant messaging is not easily accessible to researchers through an application programming interface (API). However, for technology companies and governments, data from private instant messaging offer an emerging opportunity for user profiling and targeting that seems to increasingly move into their focus (Evans, 2020; Goodin, 2021). In a similar manner to prior studies based on social media posts, this work aims to create insights into age- and gender-linked linguistic variations and explore how accurately information on user demographics can be inferred from instant messages.

1.1. Linguistic variations with age and gender

Prior studies on a variety of text sources, such as writing samples, speech transcripts, exams, or collected works of well-known writers, have investigated the association of linguistic style with age and gender in a descriptive nature (Newman et al., 2008; Pennebaker & King, 1999; Pennebaker & Stone, 2003). Findings from these studies indicate that women's language is centered around discussing people and their activities. Furthermore, women were found to use more words related to psychological processes, such as emotions (e.g., "anxious"), and social processes, for example "talk". Men's language has been identified to be rather focused on the description of external events, objects, and processes. For example, men were found to use words related to occupation (e.g., "job"), swear words, and numbers more often than women do

* Corresponding author.
E-mail addresses: timo.koch@psy.lmu.de (T.K. Koch), rp@keio.jp (P. Romero), clemens.stachl@unisg.ch (C. Stachl).

https://doi.org/10.1016/j.chb.2021.106990
Received 5 November 2020; Received in revised form 16 August 2021; Accepted 17 August 2021
Available online 18 August 2021

(Newman et al., 2008).

Regarding linguistic differences with age, research suggests that older people use more positive emotion words (e.g., "happy"), fewer negative emotion words, like "angry", and fewer self-references, such as "me". Past findings also suggest that the use of future-tense increases with older age, whereas the use of past-tense decreases and people demonstrate a general pattern of increasing cognitive complexity (Pennebaker & Stone, 2003). With the advent of computer-mediated communication (CMC), descriptive research on linguistic variations with respect to demographic differences has shifted to digital data sources, such as blogs (Argamon et al., 2007) and social media posts (Park et al., 2016), but has not yet included instant messaging data to our knowledge.

In contrast to traditional text, like books or letters, digital text is often enhanced with graphical symbols, such as emoticons and emoji. These characters are used to augment the text with additional information. Particularly in CMC, like instant messaging, emoticons and emoji play a central role. Emoticons, for example ";-)", represent facial expressions and can enrich messages with emotional context. Emoji are graphical symbols that allow giving meaning to a message, for example by adding contextual cues ("Do you want to hang out tonight? 🍺") or replacing words ("Is there still 🍕 left?") beyond the expression of emotions ("How could you do that to me? 😡"; Bai et al., 2019; Völkel et al., 2019). While user demographics play a role in the interpretation of emoji and emoticons (Butterworth et al., 2019; Herring & Dainas, 2020; Jaeger et al., 2017), research suggests that age and gender are also associated with the frequency and variety of their usage. Based on surveys (Jones et al., 2020; Pérez-Sabater, 2019; Prada et al., 2018) and real-world user data (An et al., 2018; Chen et al., 2018; Fullwood et al., 2013; Oleszkiewicz et al., 2017; Tossell et al., 2012; Wolf, 2000), researchers have found the usage of emoji and emoticons to systematically vary with demographics.

Findings on the associations of age with the usage of emoticons and emoji are diverse: Whereas an analysis of Facebook status updates suggested that younger users post more emoticons than older users do (Oleszkiewicz et al., 2017), other studies on online chat rooms (Fullwood et al., 2013) and WhatsApp messages (Pérez-Sabater, 2019) did not find significant age differences in emoticon usage. Siebenhaar (2018) analyzed the usage of emoji in WhatsApp chats and found mixed results: While he reported emoji usage to be negatively associated with age in a Swiss chat corpus, he found no age differences in an initial analysis of the chat corpus we analyzed in the present study. In a similar manner, An et al. (2018) did not find a consistent relationship of emoji usage with user age in WeChat messages. In line with theory that women experience and express emotions more often than men (Fabes & Martin, 1991; Kring & Gordon, 1998), previous research indicates that there are significant gender differences in the usage of emoji and emoticons. Findings from studies based on Facebook status updates (Oleszkiewicz et al., 2017), online chat rooms (Fullwood et al., 2013; Wolf, 2000), SMS (Tossell et al., 2012), and WhatsApp messages (Pérez-Sabater, 2019) suggested that women use more emoticons than men. Tossell et al. (2012) also found that men used a more diverse range of emoticons in their SMS data than women. The observed gender differences seem to exist for emoji, too: A large-scale study on smartphone users provided evidence that women use more emoji in their communication than men (Chen et al., 2018), contradicting a smaller study on Chinese WeChat users suggesting that gender has no effect on emoji usage (An et al., 2018). Also, women reported to use emoji (but not emoticons) more often than men in studies with self-reported survey data (Jones et al., 2020; Prada et al., 2018).

1.2. Predicting age and gender from social media posts & transfer to instant messages

Recent research on age- and gender linked language variations has extended the existing descriptive work with a prediction-oriented approach. Hereby, novel machine learning methods trained on social media text data were deployed to infer demographic characteristics of individuals or communities (Kern et al., 2016). Machine learning algorithms can be used to detect generalizable predictive patterns in rich text data sets on a large number of language features and to associate these with demographics. Using this approach, researchers were able to make inferences of users' age and gender based on language features extracted from Facebook status updates (Jaidka et al., 2018; Schwartz et al., 2013), tweets (Burger et al., 2011; Jaidka et al., 2018; Marquardt et al., 2014; D. Nguyen et al., 2021; T. Nguyen, Smith, & Rosé, 2011; Rao et al., 2010), and blog posts (Argamon et al., 2007; Marquardt et al., 2014; D. Nguyen, Smith, & Rosé, 2011; Schler et al., 2006). These studies created unprecedented insights into the associations of individual differences and language usage. By exposing how much personal information can be inferred from digital footprints on social media, this body of research also started a societal discussion about the necessity to protect individual privacy on social media.

Findings from demographic prediction studies on social media posts might not necessarily generalize to instant messages due to each channel's specific language peculiarities. For example, language usage in a given channel is also affected by its technical affordances. For instance, tweets are limited to 280 characters. Moreover, it could be shaped by the respective audience and goals of use: While private instant messaging is used to communicate with selected chat partners, social media allows to reach out to a larger readership to, for example, transmit information on one's general activities (Quan-Haase & Young, 2010). As a consequence, users engage in varying levels of self-disclosure between private instant messages and social media posts as well as across social media platforms (Bazarova & Choi, 2014; Jaidka et al., 2018). Therefore, the same user can exhibit different linguistic styles across channels (Bazarova et al., 2013; Jaidka et al., 2018). For example, prior research indicates that users prefer to self-disclose more on Facebook than on Twitter, which could be one reason[1] why language models trained on Facebook posts are more accurate at predicting users' age and gender than those trained on Twitter posts (Jaidka et al., 2018). Based on findings suggesting that users engage in more self-disclosure in private instant messages compared to social media posts (Bazarova & Choi, 2014) and reports that higher levels of self-disclosure lead to more accurate predictions of demographics (Jaidka et al., 2018), instant text messages could be more predictive of user characteristics than social media posts. As a consequence, new machine learning models trained on instant messaging data are needed in order to determine their predictiveness for user characteristics, such as demographics. However, only a limited number of small chat corpora are available to the research community yet that would allow for such predictive work (Verheijen & Stoop, 2016).

In conclusion, past findings on age- and gender-specific language variations and the successful prediction of user demographics from social media posts (e.g., Jaidka et al., 2018; Schwartz et al., 2013) motivate us to address the gap in author profiling research based on instant messages. In this work, we systematically investigate age- and gender-linked language variations in WhatsApp messages. Specifically, we replicate established methods of closed and open vocabulary approaches from existing social media research and extend our analyses to include features specific to instant messages (i.e., general message characteristics and emoji preferences). Additionally, we investigate if user demographics can be predicted from these differences in linguistic characteristics using machine learning algorithms. For this purpose, we present and apply methodological approaches to predict user characteristics even from comparably small and imbalanced text data sets. In our models, we also identify the most predictive age- and gender-related language features. Finally, we discuss implications of our findings for psycholinguistic theory and user privacy in instant messaging.

[1] Twitter's character limit could be another.

2. Method

2.1. Data set

The "What's up, Deutschland?" chat corpus was collected by Siebenhaar (2018) in Germany from November 2014 until January 2015. German WhatsApp users were invited to donate a chat conversation by exporting a WhatsApp chat of their choice as a plain text file and to email it to the researchers. Media files, like pictures or videos, were not included in the corpus due to copyright and unresolved privacy implications. We counted the placeholders from media files for quantitative analysis. Upon receipt of a chat log, an informed consent form was sent to all chat partners, stating that their text may be used and cited anonymously for scientific purposes. If the signed consent form was not returned until 14 days after the end of the data collection, the contents of all messages of the respective users were replaced by anonymous placeholders. The data of the consenting volunteers were manually anonymized: Addresses, last names, telephone numbers, location notifications, and bank account details were replaced by categorical placeholders (e.g., "Tobias" by "NAME_M" indicating a male first name). While the "What's up, Deutschland?" chat corpus is not yet available publicly, the authors kindly provided us with early access to the data.

The original corpus contains data from 495 consenting volunteers, who sent 451,938 messages in 218 chats. We excluded 260 volunteers, who did not provide demographic information on age and gender. Additionally, we removed data from nine volunteers with less than 50 words of text data, because this is the recommended minimum to obtain meaningful results from the LIWC software we used to extract linguistic features (Receptiviti, 2019). The final dataset included 162 women and 64 men of an average age of 27.08 years ($SD = 10.03$), with no substantial age difference between men (28.03 years) and women (26.70 years). The 226 volunteers contributed a total of 309,229 WhatsApp messages containing 1,968,349 words, 81,199 emoji, and 48,814 emoticons. The average volunteer submitted 1,368.27 messages ($SD = 3,391.38$) with 8,709.51 words ($SD = 20,806.01$). The volunteers used an average of 32.85 different emoji ($SD = 45.40$) and 5.74 different emoticons ($SD = 5.91$) in their donated messages. For predictive modeling, we applied a minimum threshold of 1000 words that had been used in comparable prior work on social media posts (Schwartz et al., 2013), resulting in a sample of 157 volunteers.

2.2. Language analyses

Users convey information in instant messages through a variety of means, like text, emoji and emoticons, audio files, images, or videos. Therefore, we extracted five sets of features to comprehensively quantify the characteristics of volunteers' donated WhatsApp messages (see Table 1). First, we quantified messages' text attributes through a theory-driven dictionary (LIWC), words and phrases (n-grams), and topics. This procedure represents a standard approach in language analyses of social media posts (Eichstaedt et al., in press; Kern et al., 2016). Second, we computed features quantifying general message characteristics and emoji preferences to capture additional information from instant messages. We estimated the size of gender differences for all features using Cohen's *d* effect sizes and the magnitude of the age association using pairwise Pearson correlation. We only considered coefficients where the 95% confidence interval did not contain zero.

2.2.1. Closed and open vocabulary analysis

In this work, we replicated the methods used in prior research on social media posts that made use of a combination of *closed vocabulary* and *open vocabulary* approaches in order to predict user demographics (for a detailed description of methods see Jaidka et al., 2018; Schwartz et al., 2013). Closed vocabulary methods follow a *top down* approach in form of a-priori defined dictionaries, whereas in open vocabulary methods the features are created *bottom up* from the data. Open vocabulary feature extraction methods routinely show superior predictive power over closed vocabulary approaches (Eichstaedt et al., in press).

Table 1
Computed features for the age- and gender-linked language analyses

Feature type	Number of features	Description
LIWC	96	Usage of word categories. Features were computed by the LIWC2015 software with the latest German dictionary (Meier et al., 2019).
Words and phrases (n-grams)	6,627	Single words and sequences of two to three words ("phrases") that had been used by at least 5% of volunteers. Phrases with pointwise mutual information (PMI) greater than two times the length of the phrase were kept.
Topics	2,000	Word clusters created using Latent Dirichlet Allocation (LDA).
General message characteristics	15	Length of messages; sending of media files (audio, video, and images) and contact cards; frequency of emoji/emoticon usage; range of overall emoji/emoticon usage.
Emoji preferences	179	Usage of individual emoji that had been used by at least 5% of volunteers.

Note. List of computed features from volunteers' WhatsApp messages for the age- and gender-linked language analyses.

We used the well-established Linguistic Inquiry and Word Count (LIWC) text analysis program (Pennebaker, et al., 2015) with the latest German dictionary (Meier et al., 2019). LIWC has a predefined dictionary that features words and word stems, which are categorized in theory-derived linguistic dimensions, such as standard language categories (e.g., pronouns) or psychological processes (e.g., positive and negative emotion words). LIWC counts the words in the respective word categories and computes a score for each category to indicate the relative prevalence of the words from each category in the given text. Since the word categories in LIWC are identical across languages, we can compare the scores for the word categories from our German text data with other studies based on text data in English.

Due to the absence of pre-trained topic models and age-/gender-linked lexica available for German instant messages or social media posts, as these exist for English (Sap et al., 2014; Schwartz et al., 2017), and since models are not readily transferrable across platforms and languages (Jaidka et al., 2018), we created data-driven features based on our chat corpus. Therefore, we tokenized volunteers' messages using an emoticon-aware tokenizer into single *words*. Further, we grouped the tokens into sequences of two to three words termed *phrases*. We kept phrases with a PMI (pointwise mutual information)[2] greater than 2**length*, length being the number of words contained in the respective phrase. Moreover, we kept words and phrases that were used by at least 5% of volunteers to keep the focus on common language. All word and phrase counts were normalized by each volunteer's total use of words and phrases, respectively. Similar to Schwartz et al. (2013), we extracted 2,000 *topics* by using Latent Dirichlet Allocation (LDA) with Gibbs sampling (alpha = 0.30). The LDA's underlying assumption is that documents (in our case WhatsApp messages) are a probability distribution over topics, and that topics are a probability distribution over words. In this manner, each topic is represented as a set of words with their respective probabilities. For example, one extracted topic contains the words "work", "tired", and the "-.-" emoticon, which may indicate that the sender is annoyed by work. To use topics as features, we computed the probability of a volunteer mentioning each of the 2000 topics by summing up the product of the normalized word use from that

[2] PMI quantifies the probability of the co-occurrence of words (Church & Hanks, 1990).

volunteer and the topic probability of the given word from the LDA.

2.2.2. General message characteristics

We extracted a range of features describing the general properties of volunteers' messages related to the included text, media files, contacts, emoji, and emoticons. Here, we calculated the average number of words per message, the share of messages, which contained a media file (e.g., audio, video, or image), and the share of messages containing a contact card. Further, we computed metrics on the use of emoji and emoticons. We calculated the share of messages containing any emoji or emoticon, only emoji and emoticons, the volunteers' average number of emoji and emoticons per message, and the emoji- and emoticon-to-word ratios for each volunteer. To investigate the individual range of emoji and emoticon use, we counted the number of unique emoji and emoticons used by each volunteer across all messages. We then divided this number by the total number of emoji (694) and emoticons (68) used by all volunteers in the entire corpus to express the individual ratio. In the same manner, we calculated the average range of emoji/emoticon use per message for each volunteer by dividing the number of unique emoji and emoticons per message by the total number of unique emoji and emoticons used in all messages from this volunteer. Thereafter, we divided the respective fractions by the number of messages from each volunteer. For example, if a volunteer had used 10 different emoticons in 75 messages, the relative emoticon range per message in relation to all emoticons used in the corpus was $(10/68)/75 = 0.002$.

2.2.3. Emoji preferences

In the same manner as the frequencies for words and phrases, we considered all specific emoji (179) that had been used by at least 5% of volunteers. We then counted how often each volunteer had used the respective emoji and normalized their frequency use by dividing the count by the total number of emoji used by the respective volunteer.

2.3. Predicting demographics

For the prediction of volunteers' age and gender from instant messages, we trained multiple supervised machine learning algorithms on the extracted features while accounting for the comparably small size and gender-imbalance of our data set. We compared the predictive performance of an Elastic Net (Zou & Hastie, 2005), a non-linear tree-based Random Forest (Breiman, 2001; Wright & Ziegler, 2017), and a baseline model. For the prediction of age, the baseline model would predict the mean age in the respective training set for all cases in a test set. For gender classification, it would always predict the more frequent class (in our case women) in the respective training set for all cases in a test set. We chose these particular algorithms because they allowed us to capture linear predictor effects in the data with Elastic Net models as well as non-linear effects with Random Forest models. Further, they are widely adopted in research exploring social media text using machine learning methods (Jaidka et al., 2018).

2.3.1. Model fitting on small and imbalanced data

Due to the comparably small size of our data set, we evaluated the predictive performance of our models and tuned model hyperparameter in a nested cross-validation scheme (Bischl et al., 2012). In this approach, the respective model's hyperparameters are optimized across five folds in an inner cross-validation loop, using a random search approach with ten iterations. In an outer cross-validation loop, the overall model performance with the tuned hyperparameters is evaluated across twenty folds with five repetitions. This procedure prevents an overestimation of the model's predictive performance due to model overfitting when finding optimal hyperparameters and evaluating predictive performance. For the gender classification models, the cross-validation folds were stratified in the resampling to ensure that the ratio of men to women was the same across each fold and equal to the full dataset. For Random Forest models, we tuned the hyperparameter for the number of variables available for splitting at each tree node and pragmatically set the number of trees to 1000 as a computationally feasible large number (Probst & Boulesteix, 2017). In Elastic Net models, we tuned the regularization parameter lambda and the mixing parameter alpha. Additionally, for gender classification, we used automatic tuning of the threshold values (in our case the probability value above that threshold indicates "woman", a value below indicates "man") for all algorithms. In each fold of both the inner and outer cross-validation loops, constant variables (i.e., less than 2% variance) were dropped in the process. Next, we had to further reduce the number of features since there were still way more features than volunteers. Therefore, the 1000 features with the highest Spearman rank correlation (for the age prediction) and the highest values from a Kruskal-Wallis test (for the gender prediction) in a respective training fold were retained for predictions on the test data.

2.3.2. Model evaluation

We evaluated the predictive performance of the age regression models based on the mean absolute error (*MAE*), Pearson correlation (r) between the predicted age and volunteers' self-reported age, and the coefficient of determination (R^2). In the gender classification task, we had to account for the small and imbalanced classes. In such imbalanced classification settings, it is important to consider class-specific performance metrics in addition to overall metrics. We report the prediction accuracy, F-Score (*F1*)[3], and area under the curve (*AUC*)[4] for overall performance. Further, we supply *sensitivity* (true positive rate; in our case correctly classified men) and *specificity* (true negative rate; in our case correctly classified women) to evaluate the prediction performance for both gender classes. We computed performance measures within each fold of the outer cross-validation procedure and calculated the median across all folds within each prediction model. We chose the median since it is less affected by outliers in the predictions across folds than the mean. To determine whether a model was predictive ($\alpha = 0.05$) at all, we used variance-corrected t-tests to compare the performance measures in all prediction models with those from the baseline models. These variance corrected t-tests accounted for the dependence structure of cross-validation experiments (Nadeau & Bengio, 2003). All p-values were adjusted for multiple comparisons ($n = 12$) via Holm correction.

2.3.3. Model interpretation

Because flexible machine learning models cannot be interpreted in a straight forward manner, they are sometimes referred to as "black boxes" (Yarkoni & Westfall, 2017). We used interpretable machine learning methods to increase the interpretability of our predictive models and derive relevant information for psycholinguistic theory. In order to quantify the impact of predictors in a Random Forest model trained on all predictive feature sets, we computed out-of-bag (OOB) permutation variable importance[5] (Breiman, 2001; Wright et al., 2016). For Elastic Net models, we inspected the standardized regularized regression weights to detect important variables in the predictions. Further, we created accumulated local effect (ALE) plots to visualize the effects of individual predictor variables on the predictions in Random

[3] The F−Score represents the harmonic mean of a model's precision and recall performance in one metric and ranges between 0 and 1.

[4] The AUC describes the area under the receiver operating characteristics curve when plotting the true positive rate on the y-axis against the false positive rate on the x-axis onto a two-dimensional space. AUC can range between 0 and 1, where 0 indicates the worst separability, and 1 represents a perfect separation of the classes.

[5] OOB permutation variable importance is determined by shuffling (permuting) values in the variables and by evaluating the model's prediction performance in the data that is not used for tree fitting (Wright et al., 2016). Permuting the values of unimportant variables should not affect the prediction performance, but permuting important variables should.

Forest models (Apley & Zhu, 2020). The depicted values in the ALE plots represent the mean change in predicted criterion values compared to the model's average prediction, for the given value-ranges of a predictor variable (Molnar, 2019).

2.4. Software & open materials

All data processing and statistical analyses in this work were performed with the statistical software *R* version 4.0.2 (R Core Team, 2020). For text processing, we used the *quanteda* (Benoit et al., 2018), *udpipe* (Straka & Straková, 2017), and *tm* (Feinerer et al., 2008) R packages. We extracted LDA topics using the *topicmodels* (Grün & Hornik, 2011) R package. For machine learning, we used the *mlr* framework (Bischl et al., 2016), including the *mlrCPO* (Binder et al., 2020) package for pre-processing. Further, we used the *glmnet* (Friedman et al., 2010) and *ranger* (Wright & Ziegler, 2017) packages to fit prediction models. Moreover, we created ALE plots with the *iml* package (Molnar, 2018). To make our work transparent, we provide the R code, our main figures, and results in the project's repository on the Open Science Framework (https://osf.io/ymx26/). We pre-registered our analyses before accessing the data. The pre-registration protocol and a document describing the deviations from the pre-registration protocol are provided in the repository.

3. Results

3.1. Age- and gender-linked variations

We found a range of age- and gender-linked language variations in our data. A comprehensive overview is provided in the repository. When interpreting and generalizing results, particularly of the data-driven open vocabulary features, such as words and phrases, one should keep in mind the small size of and the gender-imbalance in the data set.

3.1.1. Closed and open vocabulary analysis

Table 2 lists the top ten age- and gender-linked variations in LIWC word categories. Regarding age, words from the informal language category, particularly netspeak (including emoticons) and fillers, were written more often by younger volunteers. In the same manner, younger volunteers used first person singular (e.g., "I"), words indicating causation (e.g., "because"), and interrogatives (e.g., "why") more often. On the contrary, future-focused words (e.g., "tomorrow") and words from the family category (e.g., "children") were used more frequently by older volunteers. Moreover, the *Clout* score, which indicates high expertise, confidence, and future orientation was higher for older volunteers. Finally, older volunteers used more periods in their messages, which is closely related to the negative correlation of words per sentence with age since fewer periods suggest longer sentences to LIWC. In line with LIWC results, words and phrases indicative for informal langue, for example "ne" which translates to "a/one/no" ($r = -0.42$), "haha" ($r = -0.26$), and "geil" (engl. "hot/great"; $r = -0.26$) were used more often by younger volunteers. In the same manner, younger volunteers included more emoticons, such as ":)" ($r = -0.32$) and ":D" ($r = -0.31$), in their messages. On the contrary, older volunteers used words and phrases revolving around salutations, for example "greetings" ($r = 0.33$) or "good morning", $r = 0.29$), more frequently. Further, older volunteers used words related to work, such as "office"; $r = 0.25$), more often. Topics were not as age-discriminative and clearly interpretable as words and phrases. Hence, we refrain from interpreting these effects.

On average, women used more function words, particularly personal pronouns in first person singular, such as "I", and conjunctions (e.g., "and") than men. LIWC recognized more words from women's messages and women used more exclamation marks than men. Furthermore, women incorporated more words referring to insights (e.g., "looking forward to") and home, for example "family". On the contrary, men scored higher on the summary language variable *Analytic Thinking*, which indicates a rather formal, logical, and hierarchical thinking style in contrast to an informal, personal, here-and-now, and narrative thinking (Pennebaker, et al., 2015). Female volunteers used various forms of the verb "to go" ($d = 0.51$) more often. Men used more abbreviations, colloquial language, and words related to alcohol consumption, like "beer" ($d = -0.50$), and sex. Similar to age, gender-discriminative topics were not as clearly interpretable as words and phrases. The only distinctive female topic ($d = 0.34$) revolved around social activities, containing words such as "meeting", "seeing", and "drinking". The most distinctive male topic ($d = -0.43$) could be interpreted as salutations, containing, for example, "hey", "hi", and "xD".

Table 2
Top ten language variations in LIWC categories with volunteer age and gender

Age				
	M	*SD*	*r*	*r* CI 95%
Informal language	9.34	3.67	**−0.42**	[−0.53, −0.31]
Focus future	1.16	0.63	**0.38**	[0.27, 0.49]
(Informal language/)Netspeak	2.87	2.20	**−0.38**	[−0.49, −0.26]
Clout	58.57	17.24	**0.38**	[0.26, 0.48]
(Total function words/Total pronouns/Personal pronouns/) 1st person singular	5.19	1.77	**−0.36**	[−0.47, −0.24]
(Informal language/) Fillers	0.80	0.55	**−0.33**	[−0.44, −0.21]
(Cognitive processes/) Causation	2.53	0.84	**−0.33**	[−0.44, −0.21]
(Social processes/) Family	0.53	0.64	**0.31**	[0.18, 0.42]
(Total punctuation/) Period	7.58	5.17	**0.28**	[0.16, 0.40]
Interrogatives	1.93	0.80	**−0.28**	[−0.40, −0.15]
Gender				
	M (Men)	*M* (Women)	*d*	*d* CI 95%
Total function words	51.35	54.02	**0.67**	[0.37, 0.96]
Analytic thinking	25.80	14.55	**−0.65**	[−0.95, −0.36]
Dictionary words	83.96	86.55	**0.60**	[0.31, 0.90]
(Total function words/Total pronouns/) Personal pronouns	9.76	11.01	**0.59**	[0.29, 0.89]
(Total function words/) Total pronouns	14.80	16.10	**0.48**	[0.19, 0.78]
(Total function words/Total pronouns/Personal pronouns/) 1st person singular	4.59	5.42	**0.48**	[0.19, 0.78]
(Total function words/) Conjunctions	13.39	14.22	**0.41**	[0.12, 0.70]
(Total punctuation/) Exclamation marks	1.36	2.16	**0.40**	[0.10, 0.69]
(Cognitive processes/) Insight	1.72	1.95	**0.38**	[0.09, 0.67]
Home	0.45	0.59	**0.37**	[0.08, 0.66]

Note. $N = 226$. Table rows are ordered by absolute magnitude of the Pearson correlation coefficient for age and absolute magnitude of effect size for gender. Women are coded "1" and men are coded "0". For linguistic characteristics, the hierarchically superior LIWC categories are in parentheses. For example, the notion "(Cognitive processes/) Insight" indicates that "Insight" is a subcategory of "Cognitive processes".

3.1.2. General message characteristics

We found the usage of emoticons to be closely associated with

volunteer age. Specifically, older volunteers used emoticons less frequently and in a less diverse manner. This finding is reflected in the negative correlations of emoticon-to-word ratio ($r = -0.45$), the share of messages containing at least one emoticon ($r = -0.40$), the average number of emoticons per message ($r = -0.37$), share of messages containing only emoticons ($r = -0.19$), and the use of unique emoticons used from the entire corpus ($r = -0.18$) with age. While the frequency of emoji usage was not correlated with age, the range of emoji usage was: Older volunteers used a broader range of unique emoji overall ($r = 0.17$) and incorporated more of their own unique emoji ($r = 0.15$) in a message. Finally, older volunteers sent longer (containing more words) messages ($r = 0.20$) and more media files ($r = 0.19$).

With regard to volunteer gender, we found that women used emoji more frequently and in a more diverse manner. Specifically, women had a higher average number of emoji per message ($d = 0.54$), share of messages containing at least one emoji ($d = 0.53$), emoji-to-word ratio ($d = 0.43$), and used a broader share of unique emoji from entire corpus ($d = 0.41$) than men. For emoticons, there were no gender differences present in the data.

3.1.3. Emoji preferences

Emoji preferences varied with volunteer age. We found emoji expressing emotions, for example "🙈" ($r = -0.17$), "😂" ($r = -0.17$), "😹" ($r = -0.17$), "😹" ($r = -0.17$), and "😃" ($r = -0.16$), to be more frequently used by younger volunteers in our data set. On the contrary, we found emoji depicting objects and people, for example "🌝" ($r = 0.20$), "💋" ($r = 0.19$), "👧" ($r = 0.18$), "🌺" ($r = 0.17$), and "👶" ($r = 0.17$), to be more frequently used by older volunteers. A similar pattern emerged in the emoji preferences across genders: Women preferred emoji that express positive emotions, for example "😍" ($d = 0.51$) and "😊" ($d = 0.45$). Men, on the other hand, preferred the disappointed emoji "😞" ($d = -0.30$), representing a negative emotion.

3.2. Predicting demographics

After investigating age- and gender-linked language variations, we used cross-validated machine learning models to predict volunteers' demographics and identified predictive language features. We provide an overview of the performance of all models in the project's repository.

3.2.1. Age regression

Random Forest models ($MAE_{Md} = 3.95$, $r_{Md} = 0.81$, $R^2_{Md} = 0.49$) and Elastic Net models ($MAE_{Md} = 4.35$, $r_{Md} = 0.79$, $R^2_{Md} = 0.41$) predicted age significantly better than the baseline model ($MAE_{Md} = 6.63$, $r_{Md} = NA$, $R^2_{Md} = -0.08$). Our results suggest that all feature sets, except topics, were significantly predictive of volunteers' age above baseline (see Fig. 1). Moreover, the findings indicate that the Random Forest algorithm performed on average better than the Elastic Net algorithm in the age prediction. The large variance in prediction performance across folds for the three observed different performance measures was likely caused by the small training and test set size. For R^2, there were a few severe negative outliers, which affected the results of the respective significance tests. For example, while the variance corrected *t*-tests based on *MAE* and *r* indicated significant above-baseline prediction for LIWC features and general message characteristics, those *t*-tests based on R^2 did not.

Fig. 2 shows the most important features in the Elastic Net model (based on standardized regularized regression weights) and the Random Forest model (based on permutation feature importance) in the age prediction trained on the predictive feature sets (all features except topics). The corresponding ALE plots for the Random Forest model indicate the direction of the features' effect on the age predictions. Regularized regression weights and permutation feature importance for all predictive feature sets are provided in the project's repository. Overall, features related to the frequency of emoticon usage, specifically

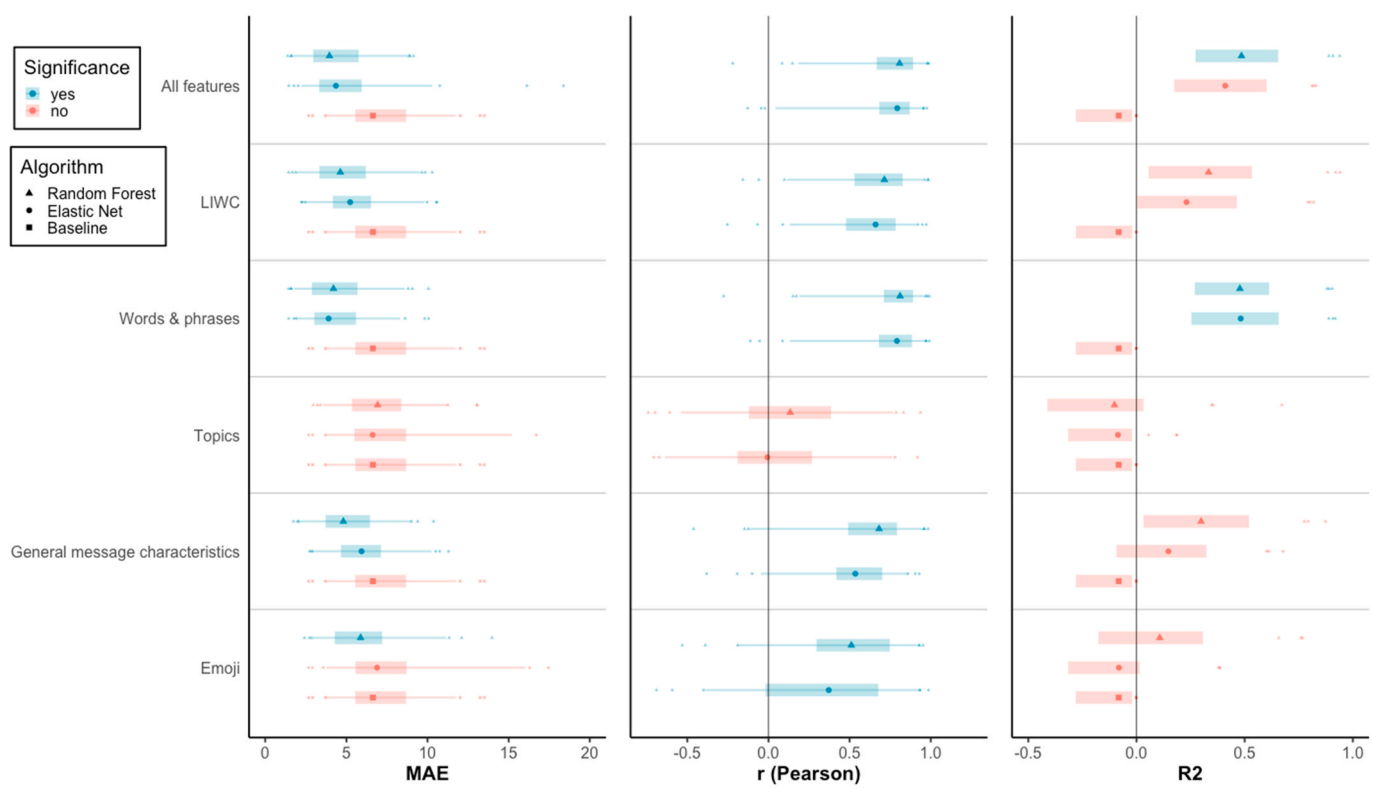

Fig. 1. Box and whisker plot of prediction performance measures from 20-fold five times repeated cross-validation for age regression for each feature (sub) set. The symbol in the boxes represents the median, boxes include values between the 25 and 75% quantiles, and whiskers extend to the 2.5 and 97.5% quantiles. Whiskers are not displayed for R^2 because of negative outliers. Pearson correlation is not available for the baseline model because it predicts a constant value, for which correlation measures are not defined. *MAE* is the mean absolute error in years. The figure is available in the project's OSF repository, under a CC-BY4.0 license.

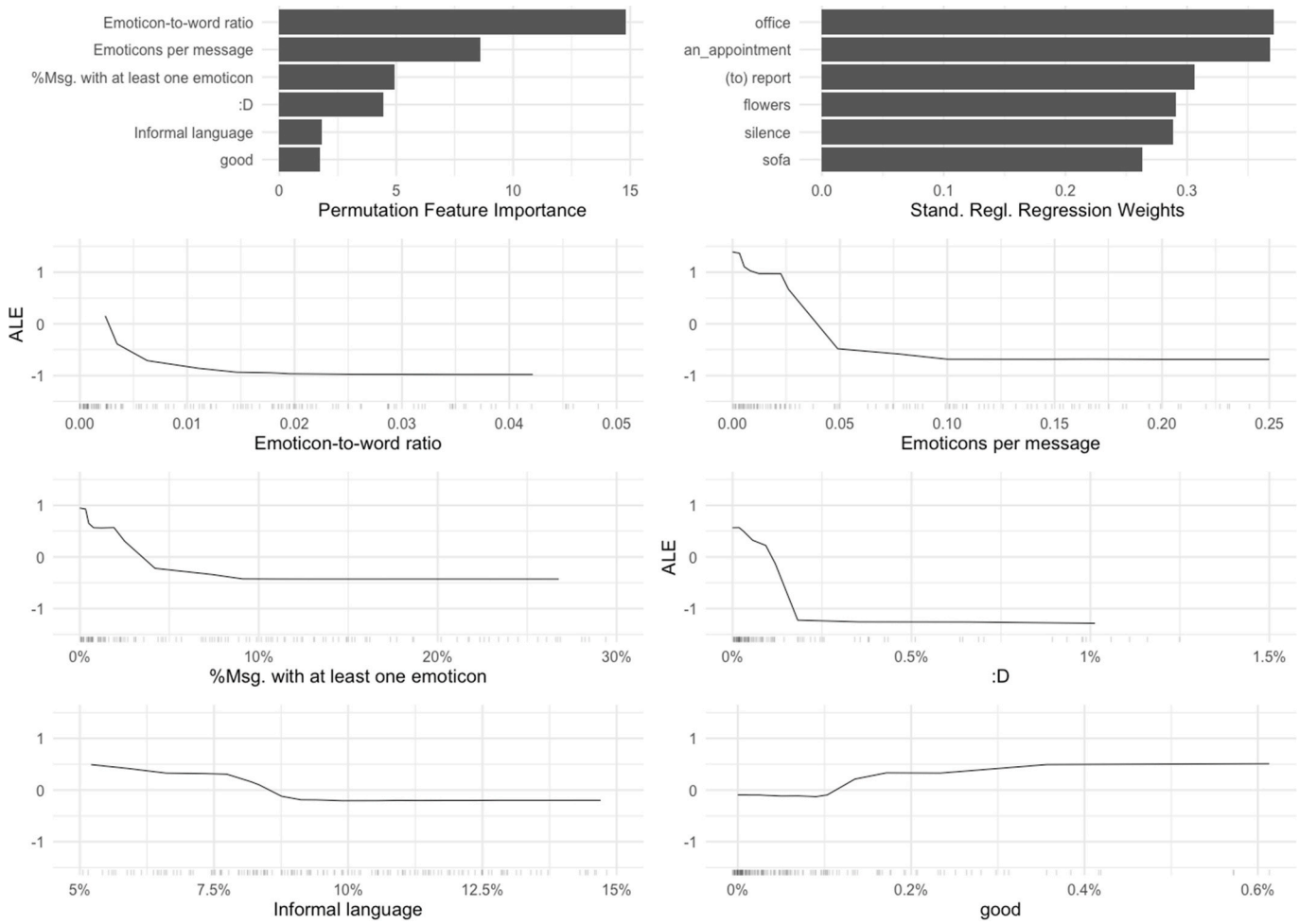

Fig. 2. Top left: Permutation feature importance for the most predictive features in the Random Forest model for age prediction. Permutation feature importance represents the decrease in the model's prediction performance (*MAE*) after permuting a single variable. Top right: Standardized regularized regression weights for the most predictive features in the Elastic Net model for age prediction. Bottom: ALE plots indicate how mean age predictions in the Random Forest model changed with regard to different values in local value-areas of the respective predictor variable. For example, the average age prediction decreases with an increasing emoticon-to-word ratio. ALE values are centered around zero. The figure is available in the project's OSF repository, under a CC-BY4.0 license.

the emoticon-to-word ratio, the average number of emoticons per message, and the share of messages containing at least one emoticon, were most important for the prediction of age in the Random Forest model.[6] This finding suggests that, for instance, if volunteers used on average less than 0.04 emoticons per message, the model predicted older age. Also, the usage of specific emoticons, such as ":D", was highly predictive in the Random Forest model, suggesting that higher usage frequencies predicted younger age. Finally, the usage of the word "good" that had often been used in salutations, such as "good morning", and the use of informal language were important for the Random Forest predictions. For example, if more than 8% of a volunteer's words were informal language, the model predicted younger age. In the Elastic Net model, the usage of words and phrases, for example "office" and "flowers", was most important.

3.2.2. Gender classification

Random Forest models ($Accuracy_{Md} = 75.0$, $F1_{Md} = 0.5$, $AUC_{Md} = 0.83$) and Elastic Net models ($Accuracy_{Md} = 75.0$, $F1_{Md} = 0.37$, $AUC_{Md} = 0.75$) predicted volunteers' gender significantly better than the baseline model ($Accuracy_{Md} = 75.0$, $F1_{Md} = 0$, $AUC_{Md} = 0.5$). Specifically, our results suggest that LIWC features and words and phrases were significantly predictive of volunteers' gender (see Fig. 3). Remarkably, models trained on LIWC features only, particularly Random Forest models ($Accuracy_{Md} = 85.7$, $F1_{Md} = 0.67$, $AUC_{Md} = 0.82$), had a higher prediction performance than models trained on all features combined. Consistent to models for age predictions, these findings indicate that the Random Forest algorithm performed on average better than the Elastic Net algorithm in the gender prediction. Also, in the same manner to the age models, there was a large variance in prediction performance across folds because of the small size of training and test sets. Further, given the class imbalance in our data, the prediction *accuracy* across folds suggests that no feature set allowed for above-baseline predictions, while *F1* and *AUC* do so. Due to the small sample size and class imbalance ($N_{\text{women}} = 114$; $N_{\text{men}} = 43$) in our data set, the models were much more accurate in classifying cases of the majority class (women) correctly than those of the minority class (men). Moreover, the models misclassified more men as women than vice versa (see confusion matrices provided in the repository). Therefore, we additionally applied the Synthetic Minority Over-Sampling Technique (SMOTE) that creates new synthetic cases for the minority class based on existing volunteers in order to balance out the training data (Chawla et al., 2002; Fernandez et al., 2018). However, applying SMOTE did not

[6] One has to keep in mind that the permutation importance scores of correlated features are ranked higher in Random Forest models. This does not indicate that they are uniquely more important for the prediction of an outcome (Strobl et al., 2008).

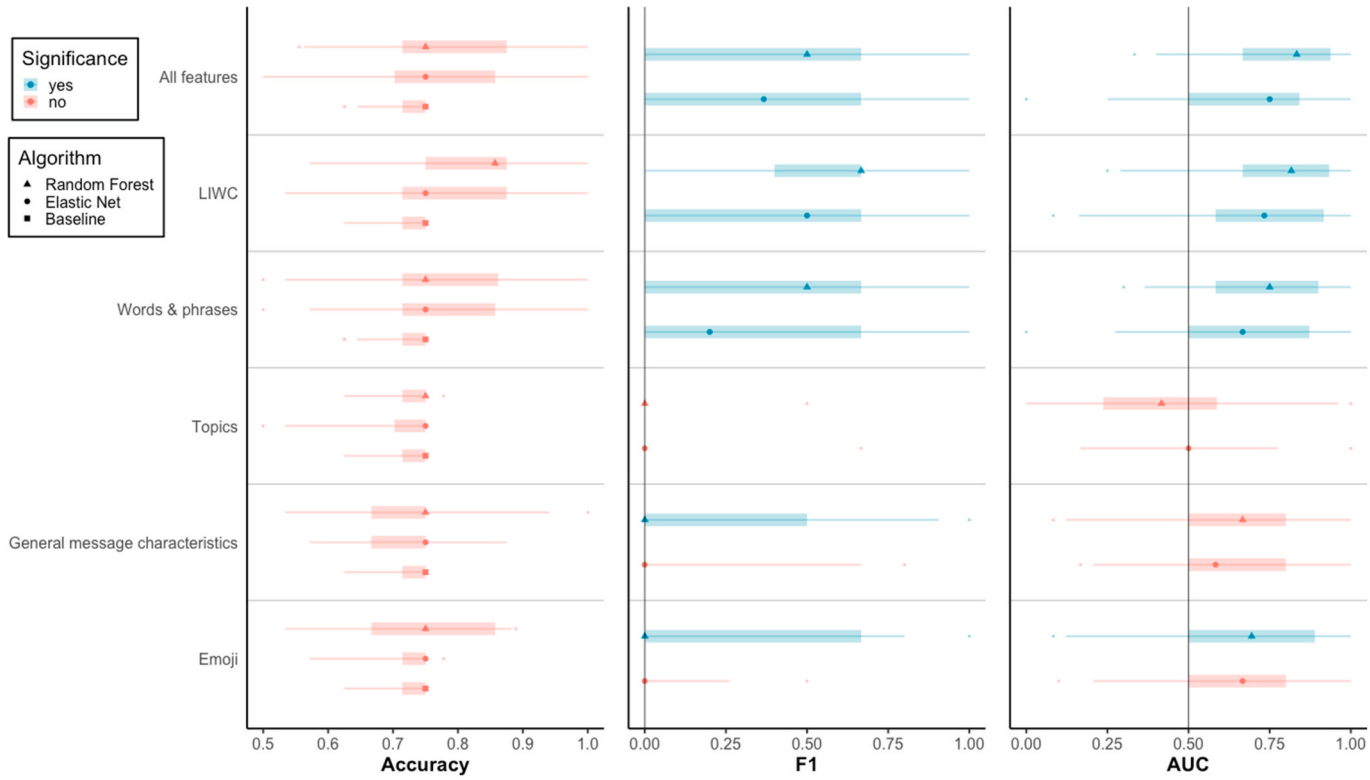

Fig. 3. Box and whisker plot of prediction performance measures from 20-fold five times repeated cross-validation for gender classification for each feature (sub) set. The middle symbol represents the median, boxes include values between the 25 and 75% quantiles, and whiskers extend to the 2.5 and 97.5% quantiles. Outliers are depicted by single points. For better readability, we omitted the baseline model because *F1* is 0 and *AUC* (Area under the curve) is 0.5 across all folds (indicated by vertical line). The figure is available in the project's OSF repository, under a CC-BY4.0 license.

improve our models' overall prediction performance over the initial approach without oversampling (see results in the repository).

Since our results suggest that Random Forest and Elastic Net models trained on LIWC features and words and phrase significantly predicted volunteer gender, we investigated features' variable importance. We refrained from including data-driven words and phrases in the feature importance analysis since the results would be specific to the small and imbalanced classes in our data set and most likely not generalize well. We found a similar pattern for LIWC features as in the descriptive gender-linked language variations: The use of function words, particularly the use of personal pronouns in first person singular, were most important with higher values, leading to a higher probability of the algorithms predicting a female volunteer. The respective figure displaying the most important features in the Elastic Net and Random Forest model (with corresponding ALE plots) is supplied in the project's repository.

4. Discussion

Our study has generated novel insights into age- and gender-linked language variations in open and closed vocabulary features, general message characteristics, and emoji preferences in instant messages. We predicted volunteer age and gender above baseline levels and identified particularly predictive features for volunteer demographics in the respective machine learning models. Finally, we presented methodological approaches to make predictions from small and imbalanced text data sets.

4.1. Age- and gender-linked language variations in instant messages

We found specific age- and gender-linked language variations to be strongly associated with volunteer demographics that were also highly predictive. For age-linked language variations, we found that younger volunteers used emoticons, first person singular, and informal language more often. Our finding that younger volunteers used emotions more frequently and that there were no age-related differences in the use of emoji usage are in line with parts of the previous literature (Oleszkiewicz et al., 2017; Siebenhaar, 2018). However, our observation that younger volunteers used emoticons more often is not in line with work by Fullwood et al. (2013), who found no variations in emoticon usage with age among users in online chat rooms. Our finding that younger volunteers used more words in first person singular than older ones is in line with results from past studies on a broad range of different text sources (D. Nguyen et al., 2021; T. Nguyen, Smith, & Rosé, 2011; Pennebaker & King, 1999; Schwartz et al., 2013). Pennebaker and Stone (2003) pointed out that this might be an indicator for people becoming less self-focused as they age. Also, our finding that younger volunteers used more informal language is in line with past studies that reported similar effects (T. Nguyen, Smith, & Rosé, 2011; Schwartz et al., 2013).

Regarding gender-linked language variations, we found that female volunteers used emoji more often, used a broader range of different emoji, and used more function words - especially first person singular pronouns. Our observation that women used emoji more often is in line with prior studies, showing that women on average use more emoji than men (Chen et al., 2018; Jones et al., 2020; Prada et al., 2018). However, this effect did not generalize to the usage of emoticons, which were almost equally often used by women and men, in our data. This finding is in line with work by Tossell et al. (2012), but does not align with results of Fullwood et al. (2013) and Rao et al. (2010), who found women to use more emoticons in their data. However, the data for those studies was collected around 2010, when emoticons were the go-to way to express emotions in computer-mediated communication, before emoji were around. Over time, the prevalence of emoticons decreased as they were gradually replaced by emoji (Pavalanathan & Eisenstein, 2015). Since our data set was collected in 2014/2015, many women possibly

used emoji instead of emoticons to express emotions, which could be the reason why gender differences in emoticon usage were not present in our data. Furthermore, women's more frequent use of function words and particularly personal pronouns in first person singular had also been found in previous studies on other text sources (Newman et al., 2008; Schwartz et al., 2013).

4.2. Predicting demographics from instant messages

Our results in the age predictions based on closed and open vocabulary features (except topics) on German instant messages compare well with previous results on English social media data that used the same methods (see Table 3). Our models performed better than prediction models in prior work by Jaidka et al. (2018) which were trained on tweets and Facebook posts in predicting user age. Schwartz et al. (2013), who trained their models on an enormous sample of Facebook posts achieved comparable performances. Many other prior studies (e.g., Marquardt et al., 2014; T. Nguyen, Smith, & Rosé, 2011; Rao et al., 2010) binned age into groups before modeling and, consequently, do not allow for a direct comparison with our results. For gender predictions, a comparison of our models' performance with prior work is not as straightforward because comparable studies only reported their overall classification accuracy (Bamman et al., 2014; Burger et al., 2011; Schwartz et al., 2013). While this metric is useful, it is highly dependent on the gender class distribution in the respective sample and makes a comparison between studies difficult. For all feature sets except LIWC features, we could not predict gender above baseline accuracy and, therefore, consider our prediction performance inferior to comparable prior work (Burger et al., 2011; Rao et al., 2010; Schwartz et al., 2013). With 85.7% accuracy for LIWC features, our models perform in a similar range of prediction accuracy as comparable LIWC models from prior work with less imbalanced samples (Jaidka et al., 2018; Schwartz et al., 2013).

Notably, even though all comparable prior research on social media text data was based on much larger data sets, the performance of our age models is similar to that of past research (Jaidka et al., 2018; Schwartz et al., 2013). A possible explanation for this observation could be that self-disclosure in private instant messages is higher compared to that in public social media posts. Consequently, instant messages could be more informative of user characteristics, like demographics, than social media posts. Future studies based on larger data sets should compare predictions from instant messages with those obtained from social media posts in a similar fashion to Jaidka et al. (2018), who compared models trained on Facebook and Twitter text.

Table 3
Predictive performance for age and gender in comparison to prior work.

Study	*N*	Data source	Features	Age: *MAE (baseline MAE)/r*	Gender: *Acc. (baseline Acc.)*
Schwartz et al. (2013)	74,859	Facebook	LIWC	−/.65	78.4 (62.0)
			N-grams	−/.83	91.4 (62.0)
			Topics	−/.80	87.5 (62.0)
Jaidka et al. (2018)	523	Facebook	LIWC	7.20 (10.06)/−	87.0 (54.0)
			N-grams	5.71 (10.06)/−	78.0 (54.0)
			Topics	6.78 (10.06)/−	91.0 (54.0)
Jaidka et al. (2018)	523	Twitter	LIWC	8.59 (10.06)/−	81.0 (45.0)
			N-grams	8.08 (10.06)/−	73.0 (45.0)
			Topics	8.58 (10.06)/−	80.0 (45.0)
Rao et al. (2010)	1,000	Twitter	N-grams	-	68.7 (50.0)
Burger et al. (2011)	184,000	Twitter	N-grams	-	75.5 (54.9)
The present study	157	WhatsApp	LIWC	4.63 (6.63)/.71	85.7 (75.0)
			N-grams	4.20 (6.63)/.81	75.0 (75.0)
			Topics	6.93 (6.63)/.13	75.0 (75.0)
			Msg. Char.	4.81 (6.63)/.68	75.0 (75.0)
			Emoji	5.87 (6.63)/.51	75.0 (75.0)
			All features	3.95 (6.63)/.81	75.0 (75.0)

Note. For comparability, we present studies using the same language features, namely LIWC, n-grams ("words & phrases"), and/ or topics. Performance measures of the best employed algorithms are reported. All prior studies were based on English text data while the text data of this work was German. *MAE* = Mean absolute error in years, *r* = Pearson correlation of predicted and true age, *Acc.* = Prediction accuracy.

4.3. Implications

By investigating age- and gender-linked language variations in instant messages, this work adds a promising new text source for the study of individual differences in language usage with relevance for psychology, computer science, linguistics and, communication research. After prior studies had demonstrated that user demographics are predictable from social media posts, our findings indicate that variation in linguistic characteristics of instant messages also allows for the accurate prediction of users' demographics, already in small samples. Such a demographic user profiling based on instant messages could be used to gain information on user demographics in order to personalize systems and for marketing efforts, based on the users' age and gender. Moreover, it could be useful to validate previously provided demographic user data. For example, this approach could be used in anonymous digital communities (e.g., only for people aged under 18) to validate user profiles and to flag suspicious profiles containing potentially false information (van de Loo et al., 2016). In order to protect users' privacy, the feature extraction and potentially model predictions should happen on the user's end, for example on the smartphone. Another field of application lies in determining the demographics of the members of an anonymous community communicating through instant messaging. By analyzing the characteristics of the messages, one could gain an approximation of the demographics of such a user population through their unique psycholinguistic characteristics. This would allow researchers to better understand the demographics of political or activist movements, and their importance to a respective populous.

These author profiling techniques have the potential for misuse, posing a threat to user's privacy and safety in instant messaging. Moreover, in contrast to public posting (e.g., on social media platforms), the design of instant messaging services does not suggest that exchanged information is accessible to third parties. Given the trend that users are increasingly shifting from social media to instant messaging, the importance of private instant messaging data is expected to rise further in the future (Goode, 2019). In this manner, Facebook has announced plans to shift their strategic focus from public posting to private messaging services (Zuckerberg, 2019) while increasing the efforts to loosen up the privacy protection of their messaging services (Goodin, 2021). Since commercial collectors of messaging data have access to much larger quantities of personal communication data, the monitoring and systematic analysis of private instant messaging environments would allow for more accurate and additional inferences beyond demographics in a similar way that social media text has been shown to be informative of, for example, personality traits and emotions (Preoţiuc-Pietro et al., 2016; Schwartz et al., 2013). The commercial accessibility of these data will likely enable timely, situation-specific targeting efforts by identifying users' momentary interests, needs, and desires in private communication. Given our findings that user

characteristics, such as their demographics, can be inferred from instant messages, even with small training samples, we argue that linguistic data from chat logs should be subject to extended privacy protection and regulation, similar to older forms of private communication (e.g., letters) or be clearly labelled as non-private.

4.4. Limitations and outlook

The results of this work are limited in three ways. First, the analyses are subject to the given data set, which is based on 226 (157 for predictive modelling) German WhatsApp users' chat language in 2014/2015, and the specific feature extraction methods we applied to the data. Predictive models from past studies on author profiling from social media text were mostly trained on large text samples with thousands of volunteers and millions of words (Schwartz et al., 2013). Our chat corpus, on the contrary, was much smaller in terms of the number of volunteers and the available text per volunteer. Further, our data was imbalanced in a way that there were more female than male volunteers and more volunteers aged 20–40 years than outside that age range. As a consequence, our models were less accurate for men and volunteers aged over 40 years because they had less data to learn from. Also, our models' age predictions got more accurate the more words per volunteer were available (see figure in repository for detailed results). More data led to more accurate and generalizable models in past work on author profiling (Eichstaedt et al., in press; Kern et al., 2016; Peersman et al., 2011). Further, our data set size is likely too small to harness the full potential of data-driven language features, particularly for topic models. Second, like many studies in the social sciences, the present work is subject to sampling biases. Specifically, the "What's up, Deutschland?" chat corpus consists of chat logs from people, who used WhatsApp, were aware of the data collection, and also decided to donate their messages for research despite potential needs for privacy. Therefore, and due to the aforementioned overrepresentation of young people and women, the data set is not representative for the general public population. Third, due to a phenomenon termed "concept drift", which describes how the underlying association between predictors and the criterion (in our case users' language usage and their demographics) changes over time (Lu et al., 2019), our results have to be interpreted in the context of the time of data collection in 2014/2015. For example, the emergence of emoji reduced the use of emoticons in recent years (Pavalanathan & Eisenstein, 2015). This trend and other developments in instant messaging have changed how men and women of varying age communicate with time. Therefore, it is necessary to retrain models on newly collected datasets. While language data from instant messaging is difficult to collect for scientific research, commercial actors would have access to larger, more representative and continuously updated samples of private instant messaging data. Hence our findings should be considered merely a conservative estimate.

We encourage researchers to replicate and extend our study with new data from a larger and more representative sample to address the limitations, and to further investigate the predictability of user characteristics from instant messages. In this context, it would be interesting to, for example, investigate whether levels of self-disclosure on public social media and private instant messaging vary across cultural and national contexts. Therefore, we would welcome if pre-trained lexica and topic models, like the ones for English social media posts (Sap et al., 2014; Schwartz et al., 2017), were made available to researchers for more languages and text sources. Moreover, while this work has exclusively focused on the message characteristics of the respective users, whose age and gender we aimed to predict, future research could also investigate the influence of the demographic characteristics of the chat partners and their message characteristics on the other person's language. It would be particularly interesting to collect additional data on the user's personality, education, language proficiency, and the relationship to the respective chat partner in order to model their language more holistically to improve prediction performance, and to evaluate the potential to infer these characteristics from instant messaging data.

The presented methodological approaches in this work can serve as guidance for future studies making predictions from small and imbalanced text data sets. Overall, to ensure comparability of studies, we want to encourage researchers to standardize methodological procedures in future prediction studies on language data. This includes treating continuous variables, such as age, in a continuous manner, reporting multiple performance measures including for the baseline model, and quantifying uncertainty in predictions by describing the variation in prediction performance across models. Finally, we suggest to investigate trained machine learning models thoroughly and apply methods to make them interpretable (Molnar, 2019; Stachl et al., 2020).

5. Conclusion

In this work, we identify age- and gender-linked language variations and demonstrate that user demographics are predictable from WhatsApp instant messages. Our findings replicate and extend past results on individual differences in social media language to the growing domain of instant messaging. Further, we provide methodological recommendations on how to make predictions from small and imbalanced text data sets. We highlight further research opportunities and emphasize the rising threats to individual privacy that could arise from the monitoring of formerly private instant messaging environments.

Supplementary material

Supplementary material is available in the project's repository on the Open Science Framework (OSF): https://osf.io/ymx26/

Credit author statement

Timo K. Koch: Conceptualization, Methodology, Formal analysis, Data curation, Writing - Original Draft, Writing - Review & Editing, Visualization. Peter Romero: Formal analysis, Writing - Review & Editing. Clemens Stachl: Conceptualization, Writing - Review & Editing, Supervision.

Declaration of competing interest

None.

Acknowledgements

We thank Prof. Beat Siebenhaar for providing us with access to the "What's up, Deutschland?" chat corpus, Dr. Florian Pargent for his statistical advice, and the members of the Psychometrics Centre at the University of Cambridge for their early-stage guidance on this project. This project was partially supported by a scholarship of the German Academic Scholarship Foundation awarded to the first author.

References

An, J., Li, T., Teng, Y., & Zhang, P. (2018). Factors influencing emoji usage in smartphone mediated communications. In G. Chowdhury, J. McLeod, V. Gillet, & P. Willett (Eds.), *Transforming digital worlds* (pp. 423–428). Springer International Publishing.

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B, 82*(4), 1059–1086.

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2007). Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday, 12*(9).

Bai, Q., Dan, Q., Mu, Z., & Yang, M. (2019). A systematic Review of emoji: Current research and future perspectives. *Frontiers in Psychology, 10*, 2221. https://doi.org/10.3389/fpsyg.2019.02221

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics, 18*(2), 135–160. https://doi.org/10.1111/josl.12080

Bazarova, N. N., & Choi, Y. H. (2014). Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network

sites. *Journal of Communication, 64*(4), 635–657. https://doi.org/10.1111/jcom.12106

Bazarova, N. N., Taft, J. G., Choi, Y. H., & Cosley, D. (2013). Managing impressions and relationships on Facebook: Self-presentational and relational concerns revealed through the analysis of language style. *Journal of Language and Social Psychology, 32*(2), 121–141. https://doi.org/10.1177/0261927X12456384

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software, 3*(30), 774. https://doi.org/10.21105/joss.00774

Binder, M., Bischl, B., Lang, M., & Kotthoff, L. (2020). *mlrCPO: Composable Preprocessing Operators and Pipelines for machine learning (0.3.6)* [Computer software] https://CRAN.R-project.org/package=mlrCPO.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., & Jones, Z. M. (2016). mlr: Machine learning in R. *Journal of Machine Learning Research, 17*(170), 1–5.

Bischl, B., Mersmann, O., Trautmann, H., & Weihs, C. (2012). Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary Computation, 20*(2), 249–275. https://doi.org/10.1162/EVCO_a_00069

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on twitter. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 1301–1309).

Butterworth, S. E., Giuliano, T. A., White, J., Cantu, L., & Fraser, K. C. (2019). Sender gender influences emoji interpretation in text messages. *Frontiers in Psychology, 10*, 784. https://doi.org/10.3389/fpsyg.2019.00784

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Chen, Z., Lu, X., Ai, W., Li, H., Mei, Q., & Liu, X. (2018). Through a gender lens: Learning usage patterns of emojis from large-scale android users. In *Proceedings of the 2018 world wide web conference* (pp. 763–772).

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16*(1), 22–29.

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., Hagan, C. A., Tobolsky, V., Smith, L. K., Buffone, A., Iwry, J., Seligman, M. E., & Ungar, L. H. (in press). Closed- and open-vocabulary approaches to text analysis: A Review, quantitative comparison, and recommendations. Psychological Methods.

Evans, D. (2020, February 22). *Why the US government is questioning WhatsApp's encryption*. CNBC. https://www.cnbc.com/2020/02/21/whatsapp-encryption-under-scrutiny-by-us-government.html.

Fabes, R. A., & Martin, C. L. (1991). Gender and age stereotypes of emotionality. *Personality and Social Psychology Bulletin, 17*(5), 532–540.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software, 25*(5). https://doi.org/10.18637/jss.v025.i05

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research, 61*, 863–905. https://doi.org/10.1613/jair.1.11192

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1.

Fullwood, C., Orchard, L. J., & Floyd, S. A. (2013). Emoticon convergence in Internet chat rooms. *Social Semiotics, 23*(5), 648–662. https://doi.org/10.1080/10350330.2012.739000

Goode, L. (2019, January 27). *Private messages Are the new (Old) social Network | WIRED. Wired.Com*. https://www.wired.com/story/private-messages-new-social-networks/.

Goodin, D. (2021, Janurary 6). WhatsApp gives users an ultimatum: Share data with Facebook or stop using the app *Ars Technica*. https://arstechnica.com/tech-policy/2021/01/whatsapp-users-must-share-their-data-with-facebook-or-stop-using-the-app/.

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13).

Herring, S. C., & Dainas, A. R. (2020). Gender and age influences on interpretation of emoji functions. *ACM Transactions on Social Computing, 3*(2), 1–26. https://doi.org/10.1145/3375629

Jaeger, S., Xia, Y., Lee, P.-Y., Hunter, D., Beresford, M., & Ares, G. (2017). Emoji questionnaires can be used with a range of population segments: Findings relating to age, gender and frequency of emoji/emoticon use. *Food Quality and Preference, 68*, 397–410. https://doi.org/10.1016/j.foodqual.2017.12.011

Jaidka, K., Guntuku, S., & Ungar, L. (2018). Facebook versus Twitter: differences in self-disclosure and trait prediction. *Proceedings of the International AAAI Conference on Web and Social Media, 12*(1).

Jones, L. L., Wurm, L. H., Norville, G. A., & Mullins, K. L. (2020). Sex differences in emoji use, familiarity, and valence. *Computers in Human Behavior, 108*, 106305. https://doi.org/10.1016/j.chb.2020.106305

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological Methods, 21*(4), 507–525. https://doi.org/10.1037/met0000091

Kring, A. M., & Gordon, A. H. (1998). Sex differences in emotion: Expression, experience, and physiology. *Journal of Personality and Social Psychology, 74*(3), 686–703.

van de Loo, J., De Pauw, G., & Daelemans, W. (2016). Text-based age and gender prediction for online safety monitoring. *International Journal of Cyber-Security and Digital Forensics, 5*(1), 46–60. https://doi.org/10.17781/P002012

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2019). Learning under concept drift: A Review. *IEEE Transactions on Knowledge and Data Engineering, 31*(12), 2346–2363. https://doi.org/10.1109/TKDE.2018.2876857

Marquardt, J., Farnadi, G., Vasudevan, G., Davalos, S., Teredesai, A., & Cock, M. D. (2014). Age and gender identification in social media. *Proceedings of CLEF 2014 Evaluation Labs, 1180*, 1129–1136.

Meier, T., Boyd, R. L., Pennebaker, J. W., Mehl, M. R., Martin, M., Wolf, M., & Horn, A. B. (2019). *"LIWC auf Deutsch": The Development, Psychometrics, and Introduction of DE-LIWC2015. PsyArXiv*. https://doi.org/10.31234/osf.io/uq8zt

Molnar, C. (2018). iml: An R package for interpretable machine learning. *Journal of Open Source Software, 3*(26), 786. https://doi.org/10.21105/joss.00786

Molnar, C. (2019). *Interpretable machine learning*. https://christophm.github.io/interpretable-ml-book/.

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning, 52*(3), 239–281. https://doi.org/10.1023/A:1024068626366

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes, 45*(3), 211–236. https://doi.org/10.1080/01638530802073712

Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2021). "How Old Do You Think I Am?" A Study of Language and Age in Twitter. *Proceedings of the International AAAI Conference on Web and Social Media, 7*(1), 439–448.

Nguyen, T., Phung, D., Adams, B., & Venkatesh, S. (2011). Prediction of age, sentiment, and connectivity from social media text. In A. Bouguettaya, M. Hauswirth, & L. Liu (Eds.), *Web information system engineering – WISE 2011* (Vol. 6997, pp. 227–240). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-24434-6_17.

Nguyen, D., Smith, N. A., & Rosé, C. P. (2011). Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities* (pp. 115–123). https://www.aclweb.org/anthology/W11-1515.

Oleszkiewicz, A., Karwowski, M., Pisanski, K., Sorokowski, P., Sobrado, B., & Sorokowska, A. (2017). Who uses emoticons? Data from 86 702 Facebook users. *Personality and Individual Differences, 119*, 289–295. https://doi.org/10.1016/j.paid.2017.07.034

Park, G., Yaden, D. B., Schwartz, H. A., Kern, M. L., Eichstaedt, J. C., Kosinski, M., Stillwell, D., Ungar, L. H., & Seligman, M. E. P. (2016). Women are warmer but No less assertive than men: Gender and language on Facebook. *PloS One, 11*(5), Article e0155885. https://doi.org/10.1371/journal.pone.0155885

Pavalanathan, U., & Eisenstein, J. (2015). *Emoticons vs. Emojis on Twitter: A causal inference approach. arXiv preprint arXiv:1510.08480*.

Peersman, C., Daelemans, W., & Van Vaerenbergh, L. (2011). Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on search and mining user-generated contents - SMUC '11* (Vol. 37). https://doi.org/10.1145/2065023.2065035

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic Inquiry and word count: LIWC 2015 [computer software]. Pennebaker Conglomerates*. Inc.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*(6), 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296

Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology, 85*(2), 291–301. https://doi.org/10.1037/0022-3514.85.2.291

Pérez-Sabater, C. (2019). Emoticons in relational writing practices on WhatsApp: Some reflections on gender. In P. Bou-Franch, & P. Garcés-Conejos Blitvich (Eds.), *Analyzing digital discourse: New insights and future directions* (pp. 163–189). Springer International Publishing. https://doi.org/10.1007/978-3-319-92663-6_6.

Prada, M., Rodrigues, D. L., Garrido, M. V., Lopes, D., Cavalheiro, B., & Gaspar, R. (2018). Motives, frequency and attitudes toward emoji and emoticon use. *Telematics and Informatics, 35*(7), 1925–1934. https://doi.org/10.1016/j.tele.2018.06.005

Preoţiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., & Shulman, E. (2016). Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 9–15). https://doi.org/10.18653/v1/W16-0404

Probst, P., & Boulesteix, A.-L. (2017). To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research, 18*(1), 6673–6690.

Quan-Haase, A., & Young, A. L. (2010). Uses and gratifications of social media: A comparison of Facebook and instant messaging. *Bulletin of Science, Technology & Society, 30*(5), 350–361. https://doi.org/10.1177/0270467610380009

R Core Team. (2020). *R: a language and environment for statistical computing*. https://www.R-project.org.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). Classifying latent user attributes in twitter. *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents - SMUC '10, 37*. https://doi.org/10.1145/1871985.1871993

Receptiviti. (2019). *LIWC API - user manual*. Receptiviti. https://www.receptiviti.com/receptiviti-api-user-manual.

Sap, M., Park, G., Eichstaedt, J., Kern, M., Stillwell, D., Kosinski, M., Ungar, L., & Schwartz, H. A. (2014). Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1146–1151). https://doi.org/10.3115/v1/D14-1121

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006). Effects of age and gender on blogging. *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, 6*, 199–205.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One, 8*(9), Article e73791. https://doi.org/10.1371/journal.pone.0073791

Schwartz, H. A., Giorgi, S., Sap, M., Crutchley, P., Ungar, L., & Eichstaedt, J. (2017). Dlatk: Differential language analysis ToolKit. In *Proceedings of the 2017 conference on empirical methods in natural language processing: System demonstrations* (pp. 55–60). https://doi.org/10.18653/v1/D17-2010

Siebenhaar, B. (2018). Funktionen von Emojis und Altersabhängigkeit ihres Gebrauchs in der Whatsapp-Kommunikation. In A. Ziegler (Ed.), *Jugendsprachen/youth languages* (pp. 749–772). De Gruyter. https://doi.org/10.1515/9783110472226-034.

Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., Gosling, S. D., & Bühner, M. (2020). Personality research and assessment in the era of machine learning. *European Journal of Personality, 34*(5), 613–631. https://doi.org/10.1002/per.2257

Straka, M., & Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies* (pp. 88–99). https://doi.org/10.18653/v1/K17-3009

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9*(1), 307. https://doi.org/10.1186/1471-2105-9-307

Tossell, C. C., Kortum, P., Shepard, C., Barg-Walkow, L. H., Rahmati, A., & Zhong, L. (2012). A longitudinal study of emoticon use in text messaging from smartphones. *Computers in Human Behavior, 28*(2), 659–663. https://doi.org/10.1016/j.chb.2011.11.012

Verheijen, L., & Stoop, W. (2016). Collecting Facebook posts and WhatsApp chats. In *Text, speech, and dialogue* (pp. 249–258). Springer.

Völkel, S. T., Buschek, D., Pranjic, J., & Hussmann, H. (2019). Understanding emoji interpretation through user personality and message context. *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI, '19*, 1–12. https://doi.org/10.1145/3338286.3340114

Wolf, A. (2000). Emotional expression online: Gender differences in emoticon use. *CyberPsychology and Behavior, 3*(5), 827–833. https://doi.org/10.1089/10949310050191809

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software, 77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Wright, M. N., Ziegler, A., & König, I. R. (2016). Do little interactions get lost in dark random forests? *BMC Bioinformatics, 17*(1), 145. https://doi.org/10.1186/s12859-016-0995-8

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B, 67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Zuckerberg, M. (2019). *A privacy-focused vision for social Networking | Facebook*. https://www.facebook.com/notes/mark-zuckerberg/a-privacy-focused-vision-for-social-networking/10156700570096634/.