Predicting social media addiction from Instagram profiles: A data mining approach

Uso del data mining per predire la dipendenza da Instagram in un campione italiano di studenti

Antonio Calcagnì, Veronica Cortellazzo, Francesca Guizzo, Paolo Girardi, Natale Canale

Abstract In this short paper, we describe an application of data mining techniques to predict Instagram users' addiction from a set of features related to (i) Instagram captions extracted from photos, videos, comments, and stories, and Instagram indicators such as number of followers and following, blocked and closed friends, and frequency of use. We first applied text mining to explore and describe the main contents of Instagram captions. Next, we used a set of non parametric models and ensemble methods to predict Instagram addiction as measured by the Instagram addiction scale [1]. Models were compared via cross-validation using test and training (random) sets from the original dataset. Results showed that Instagram addiction is mainly predicted by the overall time spent on Instagram, writing stories and comments, and number of followers. Moreover, the results suggest that Instagram users made use of photos/videos and stories/comments differently, with the latter being mostly related to emoticons, experiences, and relationships with other users.

Abstract *Questo lavoro presenta alcuni risultati di una ricerca più ampia condotta su un campione di giovani donne italiane circa l'utilizzo di Instagram come facilitatore dell'oggetivizzazione sessuale. In questo lavoro si presentano i risultati relativi allo studio circa il modo con cui attività di Instagram come pubblicazioni di foto, video, commenti e realizzazione di storie possano predire la dipendenza dal social media, come misurato dall'apposita scala di Instagram addiction. Tale*

Antonio Calcagnì,
University of Padova, e-mail: antonio.calcagni@unipd.it

Veronica Cortellazzo,
University of Padova, e-mail: veronica.cortellazzo@studenti.unipd.it

Francesca Guizzo,
University of Padova, e-mail: francesca.guizzo@unipd.it

Paolo Girardi,
University of Padova, e-mail: paolo.girardi@unipd.it

Natale Canale,
University of Padova, e-mail: natale.canale@unipd.it

analisi è stata realizzata mediante data mining, utilizzando una serie di modelli non parametrici e metodi di ensemble. La validazione e la scelta del modello migliore è stata effettuata via validazione incrociata. I risultati mettono in evidenza il ruolo del tempo trascorso su Instagram, del numero di storie, commenti e followers come elementi di predizione della dipendenza dal social network. I risultati, inoltre, hanno altresì evidenziato come gli utenti tendano ad utilizzare video e foto in maniera diversa da commenti e storie: questi ultimi, infatti, sembrano maggiormente connessi a esperienze, emozioni e relazioni rispetto ai primi.

Key words: data mining, text mining, Instagram, social media addiction

1 Introduction

Instagram is a well-known social platform commonly used for personal reasons as well as business activities. Over the last years, it has gained wide popularity across the globe, becoming one of the most popular photo-sharing applications on the Facebook platform [2]. In particular, recent trends show that Instagram is the most important network being used among adolescents [3]. Recently, a number of research have shown the role of Instagram in several psychological processes, such as women objectification (e.g., see [4]). In this respect, appearance-related comments on women's bodies accompanying Instagram images seem to play a role in body dissatisfaction as well as women self-objectification [7]. This and other results suggest the importance of investigating the interplay between social media behaviors and social media addiction, especially in young users [5].

In this short paper we will focus on Instagram addiction in a sample of Italian students. In particular, we will investigate the predictive role of Instagram contents such as text (comments, hashtags, photo captions/descriptions), indicators (number of followers/followings, blocked users, closed friends), and activity frequency on social media addiction, as measured by the *Instagram Addiction Scale* [1]. Data mining techniques have been used to analyse the data. Particularly, text mining was applied to textual components of the dataset (comments, emoticons, captions) whereas a set of non parametric models and ensemble methods has been used to choose the best model and predictors for the response variable *Instagram addiction* [6]. The results suggest that the latter is mainly predicted by activities like interactions with other users of the social networking system via messages, comments, and likes.

2 Data and Methods

2.1 Data

The *Instagram Addiction Scale* [1] was administered to $N = 97$ female participants, all of them using Instagram on a daily basis. Data were collected by using an online survey.[1] Subjects were between 18 and 31 years old with an average age of 23.64 years (standard deviation 2.23). They were asked to answer fifteen items grouped into two sub-scales (i.e., Social effect, Compulsion) using an ordinal scale with six anchors. As scales were standardized, final scores were computed by summing the items corresponding to each sub-scale. Thus, the aggregated variable *addiction* was then defined, with higher scores being indicative of higher levels of Instagram addiction. For each participant, all Instagram data were also available in compressed format (up to six months before). They consisted in *comments* to other posts, *connections* with other users (followers, following, friends, blocked users), *likes* given to media (photos and videos) as well as *comments*, *media* (photo, stories, videos), *searchers* (texts, hashtags), and *stories*. For all these data, temporal information regarding the use of the social network (time, day) were also available. The final dataset comprises 94 subjects and 95 variables, including the response variable regarding Instagram addiction. Before running the analyses, 3 subjects were excluded as they included missing observations (row-wise exclusion) whereas 36 variables were left out from the analyses because of multicollinearity (as indicated by the threshold $r \geq 0.9$).

2.2 Methods

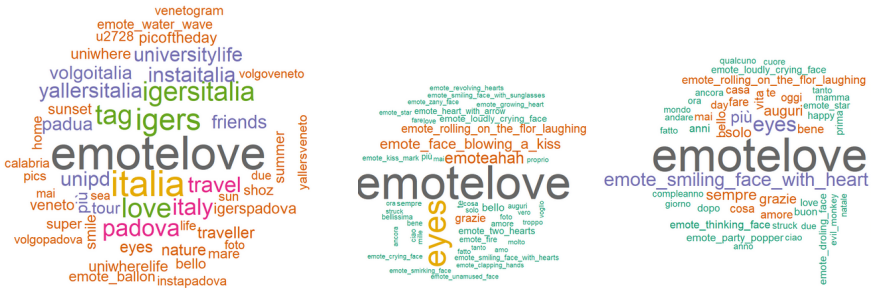
Data were analysed by means of data mining methods. With regards to the textual part of the data (photo captions, comments, hashtags), text mining descriptive techniques (i.e., most frequent terms, bigrams, graphical analyses, sentiment analysis) were used as implemented by the `R` library `TextWiller` [8] and `tidytext` [9]. All texts from Instagram were pre-processed according to standard text-mining pre-processing procedure (i.e., text-normalization, stopwords elimination, tokenization) [10]. In order to predict Instagram addiction with respect to 95 features of Instagram use, the following non-parametric and ensemble methods were instead adopted: (i) Principal component regression (tuning: number of variables), (ii) Partial least squares regression (tuning: number of variables), (iii) Lasso regression (tuning: λ penalty parameter), (iv) Regression trees (tuning: α complexity parameter), (v) Random forest (tuning: number of variables), (vi) Boosting (tuning: number

[1] The institutional review board at the University of Padova gave ethical approval for the study (protocol number: 2956)

and depth of trees), (vii) Bagging. They were compared via 10-fold cross validation, with prediction errors being computed via Out Of Bag (OOB) approach.[2]

3 Results and Discussion

Figure 1 shows wordclouds of the most frequent terms as extracted from captions of Instagram photos, videos, comments, and stories. Results indicate how distributions of photos/videos' comments were different from those of stories: indeed, words used in the first category were essentially descriptive (i.e., they just described photos/videos in terms of reference to places or experiences like traveling) whereas words used in stories/comments contained mainly emoticons and tag to users. Therefore, Instagram users made use of stories/comments and photos/videos differently. In general, sentiment analyses of captions revealed the presence of neutral sentiments more frequently than negative and positive sentiments. In particular, comments and stories showed a positive sentiment when compared to photos and videos, which suggest that users tended to use photos/videos and comments/stories with a different strategy.



(a) Photos and videos (b) Comments (c) Stories

Fig. 1: Wordcloud for captions of Photos, videos, comments, and stories.

In order to select the best subset of predictors of Instagram addiction from the large set containing 95 features, we proceeded by running a set of non-parametric models and ensemble methods as describe in section 2.2. Table 1 show the results of model comparison on training and test sets, respectively. Among others, Random

[2] More in details, ten train/test sets were created by bootstrapping the original dataset row-wise. Then, models were estimated for each training data using a 10-fold cross validation; cv-errors were computed and used to select the optimal tuned model. The best model was then chosen among those having minimum RMSE, which was computed on the test sets using the out-of-bag observations. Finally, the prediction error of the model was obtained as average of the ten RMSEs. Note that cross-validation was used as implemented in the `caret` library [11].

forest achieved lower MAE and RMSE w.r.t. parameters estimation (training set) and, in a similar way, it showed lower RMSE in the set as well. Thus, Random forest was selected as the best predictive model of Instagram addiction.

	RMSE	Test set error
PCA regression	0.7769	0.8715
PLS regression	0.663	1.0411
LASSO regression	0.7189	1.0016
Regression Tree	0.7387	0.8615
Bagging	0.6017	0.7631
Random Forest	0.6403	0.7324
Boosting	0.6749	0.7909

Table 1: Model comparison on test set. Note that Test set error is computed via bootstrap prediction error using 10 samples whereas the best model is represented in gray tones.

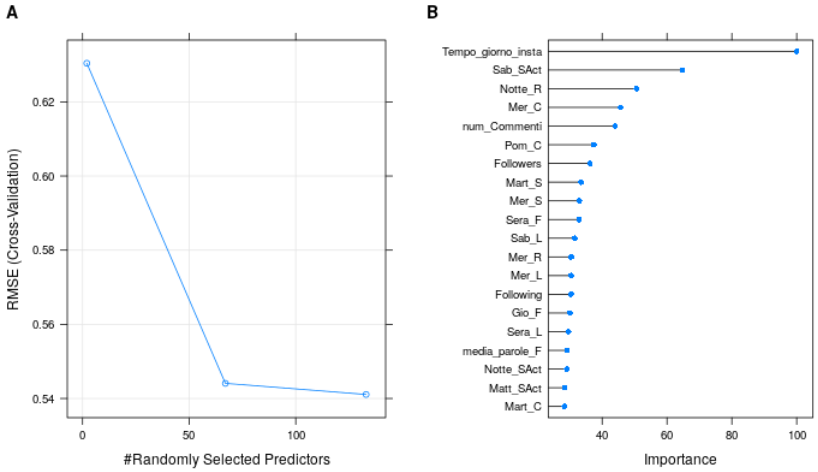


Fig. 2: Random forest to predict Instagram addiction: (A) RMSE as a function of the best number of predictors, (B) Variable importance plot for the final model. Note that variable importance measures are computed via the `importance()` function as implemented in the `randomforests` library [12].

The Random forest model was built on 500 trees which selected a final set of 67 predictors. Figure 2 shows the results of the final model. Overall, they suggested the following main predictors for Instagram addiction (in order of importance): Overall time spent on Instagram (`Tempo_giorno_insta`), replies to users' stories on Saturday (`Sab_SAct`), content searches during the night (`Notte_R`), writing com-

ments overall (`num_Commenti`) or in a specific temporal window, namely on afternoon (`Pom_cC`) and on Wednesday (`Mer_C`), number of followers (`Followers`).

4 Conclusion

We investigated the role of Instagram user's data (captions, photos, comments, likes, stories, hashtags, following and followers) in predicting social media addiction as measured by *Instagram Addiction Scale* [1] in a sample of Italian students. We used data mining techniques in order to (i) describe the use of Instagram commonly made by users and (ii) find predictors of Instagram addiction from a large database containing 95 features. The results suggested the importance of some variables, such as time spent on Instagram, writing stories, comments, and number of followers, as predictors of Instagram addiction. Moreover, the results also showed that users tend to use photos/videos and stories/comments differently: the first were mainly adopted for describing the objects they are related to. By contrast, the latter are more than simple descriptions, as they include emoticons, reference to places and experiences, tags to other users. Further studies need to be conducted to establish insights into the various mechanisms of Instagram addiction. For instance, text mining results (e.g., emoticons, sentiments) may be used to define new predictors of social media addiction, which would shed light on the role played by *typed emotions* on this emerging phenomena.

References

1. Kircaburun, K., Griffiths, M.D., Journal of behavioral addictions **7**(1), 158–170 (2018)
2. Li, Z., Agarwal, A., Management Science **63**(10), 3438–3458 (2017)
3. Brown, R.C., Fischer, T., Goldwich, A.D., Keller, F., Young, R., Plener, P.L., Psychological medicine **48**(2), 337–346 (2018)
4. Meier, E.P., Gray, J., Cyberpsychology, Behavior, and Social Networking **17**(4), 199–206 (2014)
5. Wallace, P., EMBO reports **15**(1), 12–16 (2014)
6. Azzalini, A., Scarpa, B., Springer Science & Business Media (2009)
7. Tiggemann, M., Barbato, I., Body image **27**, 61–66 (2018)
8. Solari, D., Sciandra, A., Finos, L.: Textwiller, Journal of Open Source Software **4**(41), 1256 (2019). DOI 10.21105/joss.01256. URL https://doi.org/10.21105/joss.01256
9. Silge, J., Robinson, D.: tidytext, Journal of Open Source Software **1**(3), 37 (2016)
10. Silge, J., Robinson, D.: Text mining with R: A tidy approach. " O'Reilly Media, Inc." (2017)
11. Kuhn, M., et al., Journal of statistical software **28**(5), 1–26 (2008)
12. Liaw, A., Wiener, M., R News **2**(3), 18–22 (2002). URL https://CRAN.R-project.org/doc/Rnews/

Structural entropy based modeling for psychological measurement

Modeling Strutturale basato sull'entropia per le misure in Psicologia

Enrico Ciavolino, Mario Angelelli, Paola Pasca and Omar Carlo Gioacchino Gelo

Abstract The contribution introduces the Entropy Based Structural Models and their dynamical evolution: the *Streaming entropy*. The new variant is described and tested on a typical clinical psychology scenario, that is, the psychotherapy process, as no previous work made us of behavioral data to study the psychotherapeutic relationship. Textual data consist of a psychotherapy transcript, whose word blocks have been processed and classified according to their valence: positive, negative and abstract. At first, the Stre-GCE algorithm computes one model parameter for each interaction components, then parameters get updated in an alternate way (therapist-patient, patient-therapist and so forth). Results show that Stre-GCE accounts for the fluctuating nature of the psychotherapy interaction.

Abstract *Il contributo introduce i Modelli Strutturali basati sull'Entropia ed una loro evoluzione dinamica: la Streaming Entropy. La nuova variante viene descritta e testata in uno scenario tipico della psicologia clinica, il processo psicoterapeutico, in quanto non esiste ancora un lavoro che utilizzi dati comportamentali per esaminare la relazione psicoterapeutica. I dati testuali consistono nel trascritto di un'intera psicoterapia, i cui blocchi di parole sono stati classificati in categorie: positive, negative, astratte. Inizialmente, l'algoritmo Stre-GCE stima i parametri di un modello per ciascuno dei componenti dell'interazione terapeutica, per poi aggiornarli in modo alterno con riferimento all'interlocutore precedente (terapeuta-paziente, paziente-terapeuta, ecc..). I risultati riflettono le fluttuazioni tipiche dell'interazione psicoterapeutica.*

Enrico Ciavolino
University of Salento, e-mail: enrico.ciavolino@unisalento.it

Mario Angelelli
University of Salento and INFN Lecce e-mail: mario.angelelli@unisalento.it

Paola Pasca
University of Salento e-mail: paola.pasca@unisalento.it

Omar Carlo Gioacchino Gelo
University of Salento e-mail: omar.gelo@unisalento.it