Predicting the Gullibility of Users from their Online Behaviour

Mateja Jovanović[1], Vida Groznik[1] and Marko Tkalčič[1]

[1]University of Primorska, Titov trg 4, 6000 Koper, Slovenia

Abstract

In this research we aimed to explore the predictors of gullibility in an online environment. We used machine learning algorithms to build models for predicting gullibility from social media behaviour. In total 103 Twitter users had completed the survey containing a scale for measuring gullibility. Survey data was then combined with the features extracted from the user's activity on Twitter. Besides data that was directly accessible through the Twitter API, we engineered new features containing punctuation data, usage of emojis and text vectorization with TF-IDF. This data was then standardized and reduced using Principal Component Analysis. In the modeling phase we used both regression and classification techniques. After comparison of the results with their baselines, we conclude that there is an indication that gullibility can be predicted from online behaviour. Further research and analysis are planned and are needed for a better understanding of the relationship between social media activity and gullibility. Results from this experiment showed us great potential for future work.

Keywords

gullibility, machine learning, Twitter, predictive modeling

1. Introduction

In a world filled with misinformation and people with bad intentions, gullibility has become a hot research topic. Broadly speaking, the term gullibility can be defined as "the quality of being easily deceived or tricked, and too willing to believe everything that other people say"[1]. Similarly, the definition found on Wikipedia says that "gullibility is a failure of social intelligence in which a person is easily tricked or manipulated into an ill-advised course of action"[2]. There are many different interpretations of the definition of this personal trait, however, all of the authors agree on one thing and that is the need for further research in measuring and describing gullibility. It is believed that gullibility is fully or at least partially accountable for foolish actions such as falling for romance and financial scam, political exploitation and susceptibility to fake news and other forms of disinformation [1, 2, 3, 4, 5]. Classes of people that are especially vulnerable to exploitation due to gullibility include children, the elderly, and the developmentally disabled [6].

Besides financial damage, scam victims face other problems such as trust issues and long-term

Human-Computer Interaction Slovenia 2021, November 11, 2021, Koper, Slovenia

matejajovanovicoffice@gmail.com (M. Jovanović); vida.groznik@famnit.upr.si (V. Groznik); marko.tkalcic@famnit.upr.si (M. Tkalčič)

[1]https://dictionary.cambridge.org/dictionary/english/gullibility

[2]https://en.wikipedia.org/wiki/Gullibility

trauma as a result of being a scam victim [7]. Protective organizations, banks, and insurance companies are constantly trying to inform people about threats on the internet and provide prevention systems to reduce the possibility of scams. Sadly, scammers are becoming much more creative and sophisticated with their ideas on tricking people and making a profit. Moreover, compared to the time period before the 2016 US presidential elections, there has been an increasing number of fake news. According to Google trends, people searched for the term "fake news" notably more often than before the elections[3]. In 2016, the Oxford dictionary had declared that we are living in the "post-truth" age. That term has also become the word of the year[4]. These are just some of the indicators of the power of disinformation. The impact of fake news is huge and has the potential to cause great damage in the future. Combining it with the gullibility of individuals is highly dangerous. Therefore, the goal of this work is to provide a tool for an unobtrusive detection of users' gullibility, which can help the users themselves and the agencies that wish to help the users.

2. Related work

Right from the beginning, we noticed a quite sparse set of definitions of gullibility [2, 3, 8, 5, 9, 4]. Researchers tried to address the problem of gullibility in different scenarios, mostly because of the assumptions that this trait is highly contextual. For example, Greenspan has studied gullibility in adults with intellectual disabilities [3]. He claims that this group of people is especially vulnerable to any kind of scams and is easily fooled. He claims that the accountable trait for such an unfortunate outcome is gullibility. However, adults with intellectual disabilities are just the most noticed victims of their foolish actions. The author believes that other people face gullibility as well but to a different extent and has described that the foolish action can be broken down into four parts, described in the four-factor model of gullible behavior [3]. The model is displayed in Fig 1.

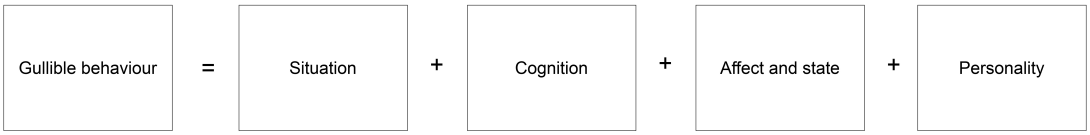

Figure 1: Gullible behaviour as modeled by Greenspan [3]

Other researchers [4, 10] have proposed that gullibility is caused by insensitivity to untrustworthiness cues. Yamagishi tested if a high level of trust is correlated with a high level of gullibility and has shown that it is quite the opposite. His results indicate that people who have higher initial levels of trust are better at detecting untrustworthiness cues and therefore less gullible than people with low initial trust levels. There is also confusion between gullibility and credulity. In his work, Greenspan has addressed this issue and made a difference between those

[3] https://trends.google.com/trends/explore?date=all&q=fake%20news
[4] https://languages.oup.com/word-of-the-year/2016/

two terms [3]. Credulity is described as a tendency to believe unlikely propositions without having supporting evidence for them. However, if those credulous beliefs involve action and there is a cause-effect relationship between them, it is defined as gullibility [3]. Taking into account the work already done in this domain, we found one research that manages to measure gullibility using a twelve item self-report gullibility scale [5]. The authors did a thorough job and performed five different studies for developing and validating their gullibility questionnaire. Moreover, this scale has been behaviourally validated in another study where participants were exposed to phishing emails[2]. Both studies showed that the 12-item gullibility scale is a reliable method for measuring gullibility. Nevertheless, even after reviewing the current state of the art methods for measuring gullibility, we were unable to find research focused on measuring the user's gullibility in an unobtrusive way, for example by using their social media activity. We believe that this could be a great benefit to understanding gullible acts in the first place, but also a useful tool for preventing potential victims from being exploited in financial, romance, political and other scams. Hence, in this paper we propose such a method.

3. Methodology

In order to devise a method for detecting gullibility from social media traces of users we used the methodology depicted in Fig 2. We first performed a pre-study to validate the questionnaire, then collected the data in the main study. We then proceeded with data pre-processing and feature engineering, finally we evaluated the predictive model.

3.1. Pre-study

Since researchers have found evidence that gullibility can be highly contextual and that it is correlated to the weak sense of self and high emotionality, we had decided to reproduce their findings. To do that we created a **pre-study** consisting of 66 items coming from 7 different scales and questionnaires. We tested the performance of this questionnaire and decided to remove some questions (the ones that do not add much information). This trade-off was made because of the long completion time and a high number of uncompleted questionnaires. The final version of the survey consisted of 42 questions and it took on average 10 minutes to complete.

3.2. Main study

Data collection has been made through a shareable link that redirected participants to the landing page hosted on 1ka.si. Participants were recruited through the personal network of the authors. Prior to filling in the survey all of the participants were given the instructions and consent form. Besides regular questions mentioned in the Sect. 3.1, we added a form where users had to input their Twitter usernames. All participants who wished to participate were asked to provide their unprotected (public) Twitter profile.

However, there were still invalid entries that we had to remove during the **data cleaning** phase. Information about user profiles and their responses to the questionnaire were stored separately in order to protect their privacy and remain their information confidential. In the

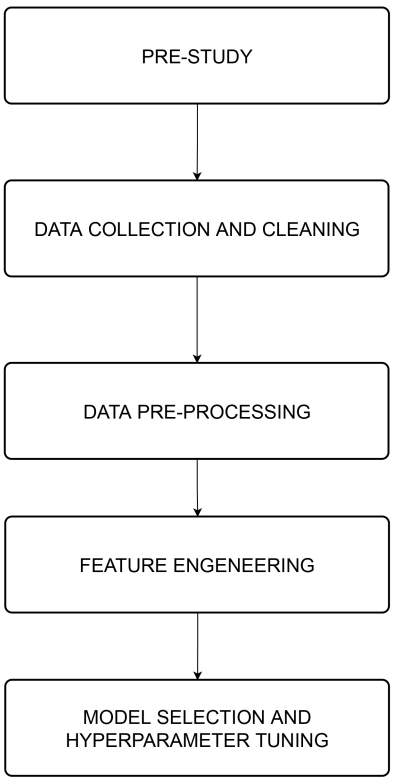

Figure 2: Methodology pipeline

data cleaning stage we had to remove all of the invalid data points. Survey entries that had been uncompleted or contained false answers to the attention check questions were excluded and considered invalid. Similarly, all of the provided Twitter profiles that were protected were excluded together with their respectful survey entries. When the data was cleaned it was time to sum up answers by each group that they are coming from, e.g. all emotionality questions were added up together to make a new variable that represented the sum of scores from emotionality questions. While we were summing up scores we were adjusting answers which had been reversely scored. Furthermore, free-form questions were converted into True/False entries. When summing up these questions we counted how many questions did user answer correctly. On the other side we were **scraping data** from their Twitter profiles. Directly from Twitter we obtained the following information: likes count, followers count, friends (followees) count, statuses (tweet, retweet, reply) count, status text (tweet's text), location, protected account (boolean), likes count received on the status, retweets count received on the status, listed count, profile's date of creation.

Nonetheless, this was not enough information to start with the modeling, therefore we started **extracting data** from the text of the acquired statuses. First, we checked the language of the statuses and created two groups of statuses per user. In one group were only statuses written in English language and in other were all statuses (including English ones). We did this because we wanted to use NLP techniques only on English language statuses, since there were participants that write statuses in two or more different languages. Inspired by researches in the use of emojis and punctuation, we decided to count all inter-punction signs, e.g. "?", "," or "-", and emojis for each user[11, 12]. For this we used the group with all statuses.

For the English group of statuses we used a common NLP approach consisting of tokenization of the text and removal of stop words. For both methods we used the NLTK library [5]. The last step was to merge all tokens of all users and do the vectorization of the words. Then we applied the TF-IDF method for giving appropriate weight to the vectors. This extracted information was then merged with the survey data and standardized using the StandardScaler[6]. Up to this point we produced a large number of input variables for the model. To reduce the unnecessary model complexity we applied the PCA (principal component analysis) dimensionality reduction technique. We have chosen 35 components to be optimal since they were explaining 69% of the variance in the data. This made our dataset ready for modeling. But before we were able to do that, we had to make sure we split the data properly. Because of our small sample size of 103 participants, we used nested five-fold cross validation for **splitting the data and hyperparameter optimization**.

In the **modeling** phase, we were predicting the variable gullibility. This variable has been made by summing up answers to the 12 questions from the gullibility scale. We used the 7 point likert scale to measure answers to each of the 12 gullibility questions. The range of the whole gullibility scale, was from 12 to 84 however, we only managed to record values ranging from 12 to 60. Additionally, we decided to approach the prediction of user gullibility both as a **regression and classification problem**. For classification models we used: random forest, gradient boosting, logistic regression, SVC and bagging in combination with SVC. For the regression models we used: SVR, ridge and stochastic gradient descent. The metrics that we used to compare the model results with their baselines were accuracy, recall, precision and f1 for classification models, and RMSE and MAE for regression models [7].

[5] https://www.nltk.org/

[6] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[7] https://scikit-learn.org/stable/modules/model_evaluation.html

4. Results

4.1. Correlation matrix

Figure 3 shows the correlations between some of the variables in the data we collected. We can observe some high correlation absolute values among variables, for example between age and cognitive reflection, between financial knowledge and financial satisfaction, and between emotionality and sense of self score. In regards to gullibility the highest correlations absolute values were detected in combination with financial knowledge and sense of self. Other features that showed a slightly lower correlation to gullibility are financial skills, emotionality, age and education.

4.2. Age-gender distribution

Figure 4 shows the distribution of genders across different age groups. The majority of the participants were between 21 and 40 years old. Out of 103 participants, 60 were males, 37 were females and six others.

4.3. Classification

In Tab. 1 we summarized the results of the classification task, where we classified each user as being either gullible or not. The baseline algorithm was predicting the most frequent class (majority classifier).

	Baseline	RF	GB	LR	SVC	Bagging + SVC
mean accuracy	0.456	0.515	0.506	0.640	0.562	0.611
std accuracy	0.036	0.106	0.068	0.128	0.098	0.096
mean precission	0.180	0.531	0.530	0.629	0.555	0.639
std precision	0.223	0.164	0.116	0.164	0.125	0.174
mean F1	0.248	0.476	0.499	0.634	0.569	0.591
std F1	0.305	0.128	0.112	0.164	0.127	0.122
mean recall	0.400	0.442	0.540	0.641	0.592	0.563
std recall	0.490	0.118	0.243	0.166	0.139	0.110

Table 1
Comparison of the classification models

4.4. Regression

In Tab. 2 we summarized the results of the regression task, where we predicted the value of the gullibility variable on the scale from 12 to 84. The baseline algorithm was predicting the average value of the predicted variable in the training set.

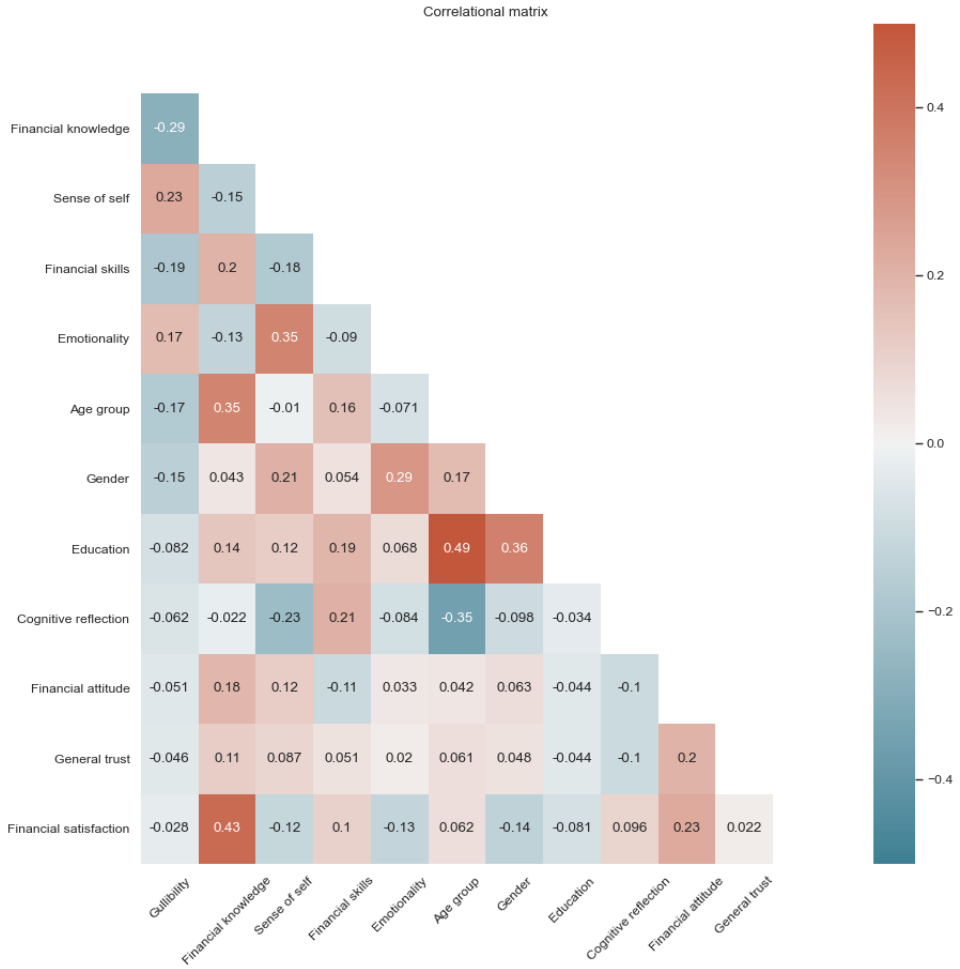

Figure 3: Gullibility correlation matrix

	Baseline	SVR	Ridge	SGD
mean RMSE	10.345	10.794	10.002	10.036
std RMSE	1.654	1.662	1.646	1.640
mean MAE	8.600	8.675	8.201	8.183
std MAE	1.458	1.288	1.343	1.295

Table 2
Comparison of the regression models

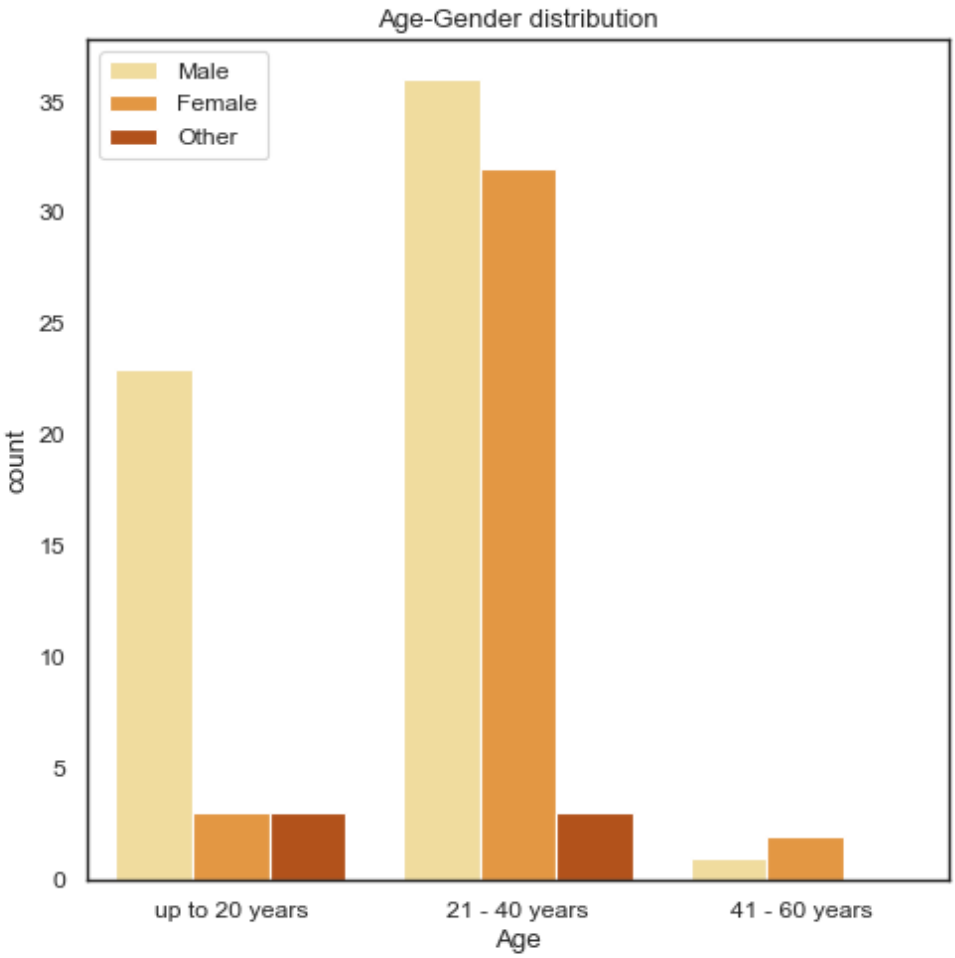

Figure 4: Age-gender distribution

5. Discussion

Results from the figure 3 showed that gullibility is negatively correlated to the financial knowledge and financial skills which are part of the financial literacy questionnaire. All of the correlations values from the matrix were generally low but, in comparison to the other features financial knowledge and financial skills have a high absolute correlation to gullibility. This is important because we added the financial literacy questionnaire to our survey in order to investigate if there is a relationship between gullibility and this fairly contextual feature. These findings are not enough to support any claims about gullibility, but they represent the first step

towards new findings in this direction. Besides mentioned features, sense of self, emotionality, age and gender were showing signs that they are correlated gullibility. Emotionality and age were expected to be in this group since we know that other researchers had similar results [2, 6]. Surprisingly, sense of self was positively correlated to gullibility, even though other evidence shows that a weak sense of self is correlated to gullibility[2].

After the comparison of the model's performance we can say that both approaches, classification and regression, performed better than their baselines. We reported the average results from 5 different splits to make results more reliable and avoid optimistic bias caused by lucky split. In the Tab. 1 we can see the results of the classification models compared next to the baseline results. The best performing classification model was a logistic regression with a mean accuracy of 0.640 however, when interpreting the average result we should take into account the standard deviation. Logistic regression also had the highest standard deviation (0.128) from all classification models. If we take a look into precision metrics we can see that Bagging in combination with SVC performed slightly better than the logistic regression. The Tab. 2 represents the results of the regression models compared to their baseline results. The baseline was calculated by taking the average result from all splits. Results did not vary much across the models. The only model that underperformed and had worse results than the baseline was the SVR model.

6. Limitations and future work

Possible limitations of this research could be the small sample size. We have planned to extend our research in order to solve this issue and gain statistical significance over our results. Also, there is a possibility that highly gullible people are not using Twitter, for example elderly people[6]. Besides this we believe that the models have shown any indication that gullibility can be measured from users' online behaviour. In our further research on this topic we will try to use more sophisticated language models, that would enable us to utilize the information from non-english tweets as well. We have tested if gullibility is correlated with financial literacy and failed to report a statistically significant correlation. This could be due to the complexity of the questions used to measure financial literacy. However, the correlations between financial knowledge and gullibility and financial skills and gullibility were in the top three highest correlations in respect to gullibility. For future work we also suggest testing out different (simpler) questionnaires for financial literacy.

References

[1] The social psychology of gullibility: Conspiracy theories, fake news and irrational beliefs, Routledge, 2019.

[2] M. S. George, A. K. Teunisse, T. I. Case, Gotcha! Behavioural validation of the Gullibility Scale, Personality and Individual Differences 162 (2020) 110034. URL: https://doi.org/10.1016/j.paid.2020.110034. doi:10.1016/j.paid.2020.110034.

[3] S. Greenspan, Chapter 5 Foolish Action in Adults with Intellectual Disabilities. The Forgotten Problem of Risk-Unawareness, volume 36, 1 ed., Elesvier Inc., 2008. URL: http://dx.doi.org/10.1016/S0074-7750(08)00005-0. doi:10.1016/S0074-7750(08)00005-0.

[4] T. Yamagishi, M. Kikuchi, M. Kosugi, Trust, gullibility, and social intelligence, Asian Journal of Social Psychology 2 (1999) 145–161. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-839X.00030. doi:https://doi.org/10.1111/1467-839X.00030.

[5] A. K. Teunisse, T. I. Case, J. Fitness, N. Sweller, I Should Have Known Better: Development of a Self-Report Measure of Gullibility, Personality and Social Psychology Bulletin 46 (2020) 408–423. doi:10.1177/0146167219858641.

[6] S. Greenspan, Annals of gullibility: Why we get duped and how to avoid it, Praeger, 2008.

[7] D. Glodstein, S. Glodstein, J. Fornaro, Fraud trauma syndrome: The victims of the bernard madoff scandal., Journal of Forensic Studies in Accounting & Business 2 (2010).

[8] H. Mercier, How gullible are we? A review of the evidence from psychology and social science, Review of General Psychology 21 (2017) 103–122. doi:10.1037/gpr0000111.

[9] S. Greenspan, G. Loughlin, R. S. Black, Credulity and gullibility in people with developmental disorders: A framework for future research, International Review of Research in Mental Retardation 24 (2001) 101–135. doi:10.1016/s0074-7750(01)80007-0.

[10] J. B. Rotter, Interpersonal trust, trustworthiness, and gullibility., American Psychologist 35 (1980) 1–7. URL: https://doi.org/10.1037/0003-066x.35.1.1. doi:10.1037/0003-066x.35.1.1.

[11] A. A. Md Shoeb, S. Raji, G. De Melo, Emotag - Towards an emotion-based analysis of emojis, International Conference Recent Advances in Natural Language Processing, RANLP 2019-September (2019) 1094–1103. doi:10.26615/978-954-452-056-4_126.

[12] D. N. Gunraj, A. M. Drumm-Hewitt, E. M. Dashow, S. S. N. Upadhyay, C. M. Klin, Texting insincerely: The role of the period in text messaging, Computers in Human Behavior 55 (2016) 1067–1075. URL: http://dx.doi.org/10.1016/j.chb.2015.11.003. doi:10.1016/j.chb.2015.11.003.