Coursework - Data Science

CID: 02091191

April 28, 2022

Abstract

Individuals, media, governments and NGOs are present on social networks and play a key role in sharing information. Relaying false information has negative impacts on society, such as manipulating public opinion. An online survey provided demographic data on more than 2,500 participants and their behaviour when confronted with disinformation messages in different situations. Encouragingly, the data show that very few people are likely to share disinformation on social networks. It is also possible to see that men and women are not affected to the same extent by this phenomenon. Furthermore, the data show that the majority of people share fake news on social networks because they believe in the messages, that it confirms their initial thoughts on a subject (usually political) and that they have seen it before. To a lesser but still important extent, factors such as age, education, gender or social conservatism matter more than the reliability of the source, the consensus around the message or even personality traits. The aim of this study is to design a procedure which takes into account predictors for likelihood of sharing together and study their importance in presence of other variables and not only the single effect of each feature on the likelihood of sharing.

Introduction

It would be absurd to deny that social networks have become such a huge part of our society that traditional media such as television or newspapers have seen their audience gradually decline in recent years. But are social networks reliable sources? Probably not, at least not always. It has therefore become essential to recognise what is true and what is false in the information that is relayed on them. Fake news can take two forms, which are misinformation and disinformation [1]. Roughly speaking, disinformation is a subset of misinformation which is deliberately misleading. While the majority of the population is concerned about the problems of misinformation on social networks [2], only a few are able to tackle the issue. The dissemination of fake news on social networks poses problems of information manipulation for a variety of purposes, such as manipulating opinion on a conflict or an election. This was the case, for example, for the war in Libya, which was very well publicised on social networks, where you could see buildings destroyed by rockets and where each side accuses the other of being responsible. Another example is the US presidential election in 2016 when Max Read, an editorial writer for New York Magazine, wrote an article entitled *Donald Trump won because of Facebook* the day after the election and gave rise to the Cambridge Analytica scandal. Hence, disinformation on social media is at the centre of nowadays stakes.

The considered dataset was collected on the basis of four online surveys, which aimed to understand the role of certain variables in modelling the likelihood of sharing misinformation on social networks. This likelihood is measured by the participants themselves according to their likelihood of sharing misinformation messages shown to them. The four studies are similar in terms of variables, with the only difference being that study 1 involves Facebook users, study 2 involves Twitter users and study 3 involves Instagram users all in the UK and study 4 involves Facebook users in the US. The full methodology can be found in the technical report [3]. This study makes sense in the light of recent moves by the digital giants to better regulate their platforms, which is not an easy task. Indeed, it seems a priori consistent that demographic variables and others related to the messages that can be posted allow to better understand the phenomenon. Thus, by better understanding the mechanisms behind the sharing of fake news, platforms could better fight against it [4].

1 Background

Various papers in the literature focus on factors influencing the sharing of disinformation on social networks. Literature has divided features of interest within two groups: message characteristics and recipient characteristics [5]. It seems consistent to assess that dissemination of fake news depends on both factors. A title can be exaggerated and teasing, an illustrating image can be false and out of context, the source is more or less reliable, the message is more or less relayed: these are examples of message characteristics. Age, education, political beliefs, gender are examples of recipient characteristics.

One may first think that person characteristics are important, as in [6] which focuses only on demographic variables in order to better understand the spread of disinformation on Facebook. Similarly to [7] and [5], this paper based on self-reported surveys conducted during the 2016 US elections shows that only a few people share disinformation (or equivalently fake news) on Facebook. This part of social network users is estimated to be less than 10% in [5]. While many recipient features seem to help modelling the spread of fake news on Facebook, some are statistically more significant than others. Using a Poisson model, [6] finds that political beliefs, age and gender are statistically more significant than ethnicity, education or income to model the likelihood of sharing disinformation, measured as the number of fake news articles shared.

One limit of this study is the few numbers of features, in addition that these features are only focused on demography, in other terms on people sharing messages and not messages themselves. Other papers such as [7] suggest considering personality traits to better understand mechanisms behind sharing disinformation online. Taking into account personality traits (explained `here`) along with age, gender and trust in the shown information, this paper shows that among all personality traits, only conscientiousness was associated with participants' ratings of their likelihood to engage with the material. The more conscientious, the less likely to share disinformation, which is consistent since more conscientious and careful individuals are checking what they share on social networks. Moreover, trust in the piece of information is statistically significant in their models and tends to increase the likelihood of sharing disinformation on social networks. This is a long established idea according to which people are more likely to share content in which they believe and that agree their preexisting beliefs. However, gender does not seem to be statistically significant in their model which is in disagreement with [6].

While some features are reported to be significant in a paper, the same features could be said as irrelevant in other studies. In fact, it depends on multiple aspects: data used in all papers are not the same and features neither. Statistical significance must be interpreted in the presence of other variables, which could explain why gender seems significant in [6] but not in [7]. To focus more on gender differences in sharing disinformation on social networks, [2] suggests to focus only on gender as a feature to model the concerns, the detecting and actual topics between genders about the stakes related to this problem. It is shown that both men and women face difficulty to identity fake news in a similar way, although women are more concerned than men about the effects of disinformation on society. Relationships between concern and detecting fake news is assessed: the less concerned about disinformation are the same people who claim to have fewer difficulties in detecting it. However the study does not find any significant difference between men and women in detecting fake news, it shows that men are more likely to share disinformation content about sports and politics than women, whereas women are more likely to share disinformation content about celebrities.

A few studies have tried to focus on both message and recipient characteristics such as [5] which uses the same dataset as in our study and tries to find both recipient and message factors that make people share disinformation. Trusting in the message increases the likelihood of sharing disinformation: this makes sense regarding the "activating the base" principle according to which people share content in which they trust [7]. Familiarity with the material also increases the likelihood of sharing since it is a sign of the willingness to give more visibility. Personality traits like low agreeableness and low conscientiousness come with increasing likelihood of sharing as well as in [7], but it depends on the platform. Some features that we could think at first sight relevant can turn out to be less significant than we thought, such as demographic variables (gender, age), authority (the reliability of the source, its number of followers) or the consensus about the information (the number of likes/retweets/share). More surprisingly, this study finds that higher digital media literacy might actually be associated with a higher likelihood of sharing false material. This study considered each social media (Twitter, Facebook, Instagram) completely separately, which is acceptable since people behaviours are different on each network, but is also a drawback since it multiplies the number of relations with each feature to the likelihood of sharing disinformation online, makes the study more exhaustive and less readable. Also, the author seems not to be clear about its definition of *relation* between a feature and the response, which may be interpreted as a correlation measure.

2 Data and methods

2.1 Description of the data

The considered dataset [3] in our study is the result of a survey commissioned by the Centre for Research and Evidence on Security Threats (CREST) in 2019, which aim is to explore some of the potential factors influencing the spread of disinformation on social media. Four different online experimental studies were conducted with messages characteristics being manipulated in order to collect data about individuals. In fact, the data collection was split into four scenarios focusing on different social media (Twitter, Instagram and Facebook in the US and the UK) and where 2,634 participants self-reported the effects of message attributes and their viewer characteristics on the likelihood of spreading disinformation. The psychometric measures used were the New Media Literacy Scale (Koc & Barut, 2016), the Social and Economic Conservatism Scale (Everett, 2013), and a Five-Factor personality questionnaire (Buchanan, Johnson, & Goldberg, 2005) derived from the International Personality Item Pool (IPIP; Goldberg, 1999). Message attributes include authoritativeness of the source, consensus measurements and consistency with pre-existing beliefs. Viewer characteristics include personality, demographic data and digital media literacy. To measure their likelihood of sharing disinformation online, participants were shown examples of false information found on social media and rated how likely they would share it on their social networks. In addition to their likelihood of sharing, they were asked to rate how likely these messages could be true and how likely they thought they had seen this information before.

Most pieces of information can be gathered in order to obtain scores on conservatism or on personality for example, using existing methodology as explained previously. Therefore, using a subset of all variables is acceptable. The amount of missing data in the dataset is relatively small and there is no particular pattern in missing data: participants just did not mention the answer to some questions such as their age, personality traits or digital media literacy. In fact, simply deleting samples where there are missing data would represent a 2% loss of data which may be considered as negligible. Thus, rows with missing data are discarded in this study and the sample size becomes $n = 2579$. Considered variables are:

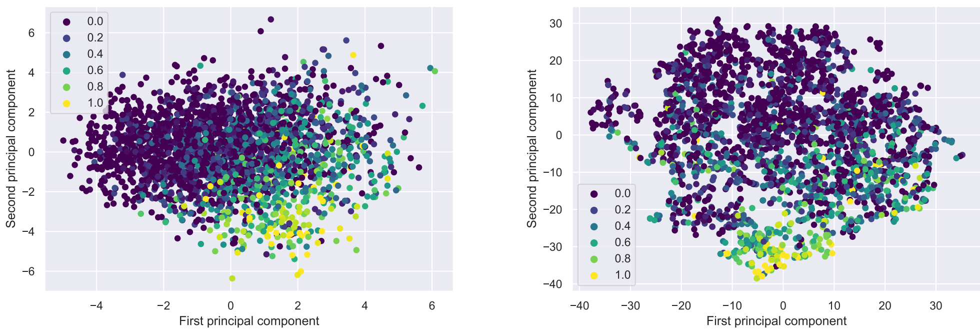
- **Country**: categorical, U.S. or U.K.
- **Age**: continuous.
- **Gender**: categorical, the proportion of gathered "Other" or "PNTS" samples is less than 10, which is not enough to draw any significant conclusion about this gender. Considering only male and female, a simple column is sufficient (1 for female, 0 otherwise).
- **Education**: categorical, was encoded according to the number of years of studies for each answer.
- **Employed**: categorical, given the occupation, we split participants in employed (self-employed or for wages) and unemployed (students, looking for work).
- **Right**, **Left**: categorical, according to political beliefs. Centre is considered but discarded as a column to avoid co-linearity since information about Centre is already contained in Left and Right (if Left and Right are 0, then Centre is 1).
- **Frequency of use**: categorical.

- **Shared (found later), shared (while knowing)**: categorical, participants who already have shared false information on social networks and realised after it was false (found later) or knew it was false before posting (while knowing).
- **Authoritative**, **Consensus**: categorical, reaction of participants if the message source is famous (authoritative) or if it is highly shared (consensus)
- **Openness**, **Neuroticism**, **Extraversion**, **Conscientiousness**, **Agreeableness**: continuous, obtained scores from personality test.
- **Social Conservatism**, **Economic Conservatism**: continuous, scores measuring social and economic conservatism.
- **Functional and Critical Consumption and Prosumption**: continuous, measures of digital media literacy.
- **Facebook**, **Twitter**: categorical, corresponds to the survey the participant answered to. Instagram could be discarded as a column since the information is already contained into Facebook and Twitter columns.
- **Likelihood sharing, trust and seen before**: continuous, likelihood of sharing, trusting and having seen before disinformation content on social networks.

Thus, the initial dataset can be converted to a more suitable object for data exploration and modelling since it is now numerical. Since all features have different ranges of values, all variables are standardised. The only exception to that transformation is the likelihood of sharing which is simply re-scaled between 0 and 1. Note that these transformations are made to improve future models, and have no influence on data exploration. Thus, the transformations were performed after the following data exploration.

2.2 Exploratory analysis

Since this dataset is high dimensional, one could reduce dimension to help visualising the data.



(a) PCA (b) t-SNE

Figure 2.1: Low dimension visualisation of the features space using two dimensionality reduction techniques. Colour scale corresponds to the scaled likelihood of sharing disinformation.

Figure 2.1 shows representations of the features space in two dimensions using principal component analysis (PCA) and t-SNE algorithm. Note that the features used here did not take into account the target variable which is the likelihood of sharing fake news on social networks. However, these plots show that there may be some patterns within the features to model the likelihood of sharing disinformation since the obtained manifolds seem to have organised data points with high likelihood of sharing together. Therefore, trying to build models for the response variable from the considered features seems consistent.

At first sight, some features could appear more important than others when considered uniquely such as frequency of use, age or political beliefs. Figure 2.2 shows that people who use the most social networks are overall more sharing disinformation on social networks than rare users. Dark bars represent 95% confidence interval created using 1000 bootstrap samples: these bars are sometimes wide since there is not a significant part of old and frequent users in the dataset for example. Using age to model directly the likelihood of sharing would not be very useful, although it seems that likelihood of sharing increases with age except for the youngest and the oldest samples. Figure 2.3 shows the likelihood of sharing fake news of respondents who identified as left, right, or centre politically oriented. This illustrates first the small number of people who are likely to share fake news on social networks, and it then shows that the proportion of people voting right increases with likelihood of sharing disinformation which makes sense since the fake news used in the survey were right-oriented.

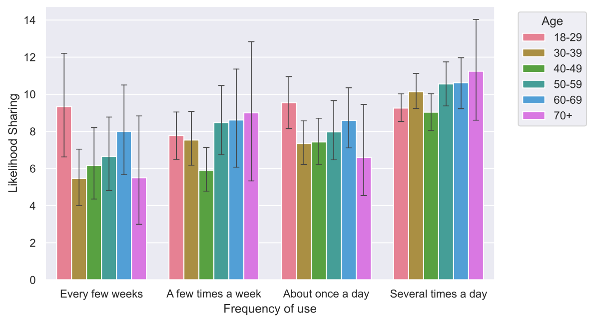


Figure 2.2: Bar plot of likelihood of sharing according to frequency of use and age group.

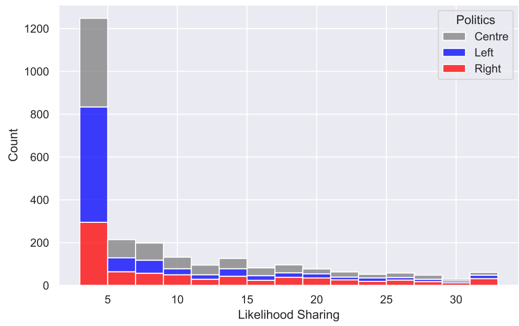


Figure 2.3: Histogram of likelihood of sharing according to political beliefs.

One could also take a glance at correlation between all variables to look for redundancy as shown on Figure 2.4. Some features are more correlated to the likelihood of sharing disinformation online such as the likelihood of trust in these messages or to have seen them before, voting right, age (the older the more likely) and use Facebook. Some features like personality traits are less correlated to the target variable, and others are negatively correlated with likelihood of sharing disinformation online such as openness or voting left.

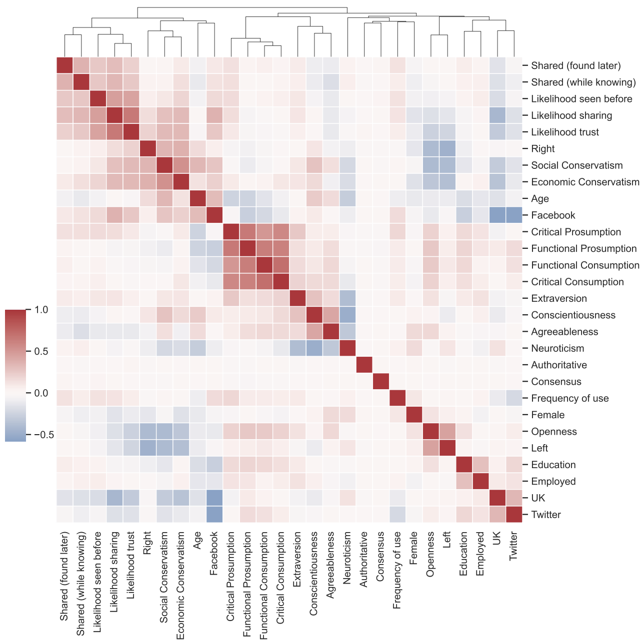


Figure 2.4: Correlation matrix heatmap between variables. A hierarchical clustering algorithm is used to gather features by similarities and its dendrogram is drawn on top.

Finally, social and economic conservatism could help modelling likelihood of sharing disinformation online. Figure 2.5 shows that likelihood of sharing seems to increase with both social and economic conservatism. Facebook may be the most used social network for spreading disinformation online according to the survey since Facebook points are overall bigger and Instagram does not seem to be much impacted by disinformation. It does not mean that Instagram is free of disinformation and one may suppose that other topics of disinformation can spread on Instagram, such as celebrities.

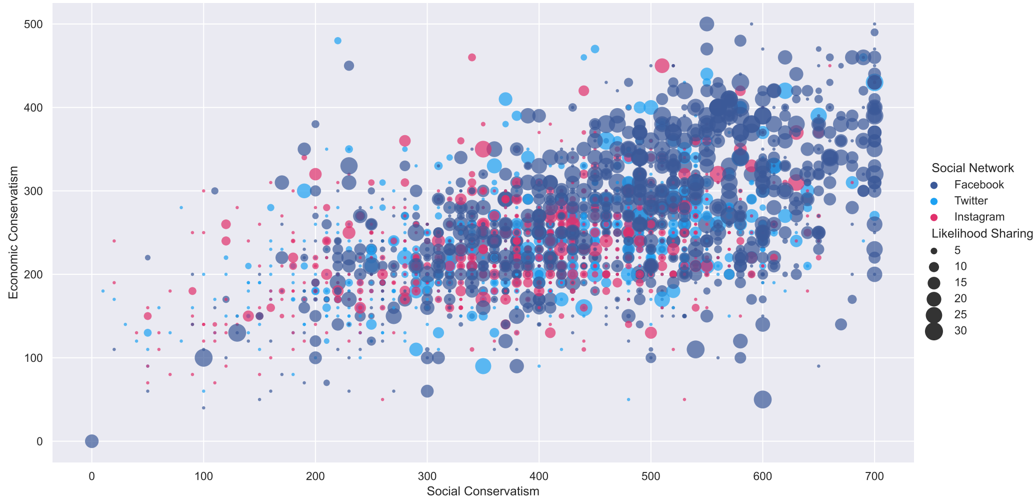


Figure 2.5: Scatter plot of economic conservatism against social conservatism. Size represents the likelihood of sharing and colour represents the social network used.

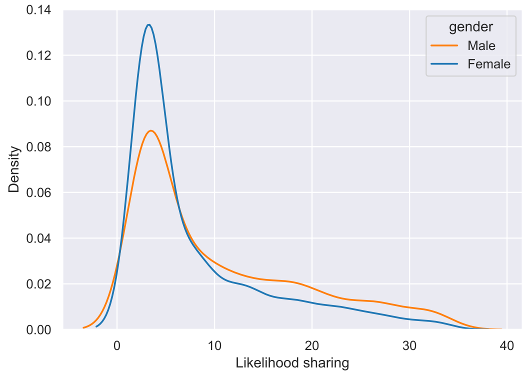
3 Results

3.1 Relationship between likelihood of sharing and gender

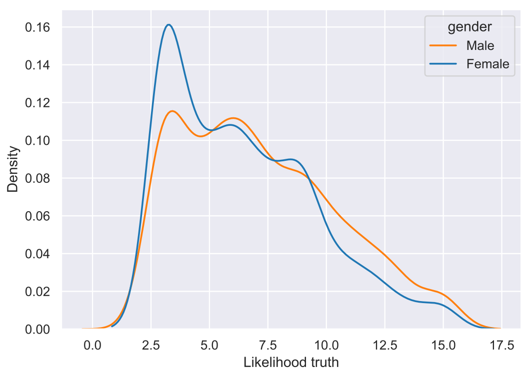
Assume in this part we try to model the likelihood of sharing disinformation only using gender information about participants and no other variables. Recall that initially the dataset contained Other and PNTS genders which sample size was not sufficient to draw any significant conclusion on them. Thus, we consider here only male and female genders. The data contains 1115 male samples and 1455 female samples.

A first idea would be to compare the two sample means likelihood of sharing between genders using hypothesis testing, in particular a t-test. The two hypothesis are the null hypothesis H_0: $\mu_1 = \mu_2$ and the alternative hypothesis H_1: $\mu_1 \neq \mu_2$ where μ denotes the mean likelihood of sharing disinformation per gender and 1, 2 denotes respectively male and female. The test statistic t follows a Student t-distribution with 2157 degrees of freedom, and we find numerically $t = 7.586$ which is associated with a $4.882 \cdot 10^{-14}$ p-value. Therefore, there is enough evidence to reject the null hypothesis which means that according to collected samples, men and women have significantly different likelihood of sharing disinformation on social networks. Since $\mu_1 = 7.153 > 6.469 = \mu_2$, we could conclude that according to this test men are more likely to spread disinformation on social networks than women.

This result is also confirmed using a normal linear model when modelling the likelihood of sharing disinformation online according to gender groups using Helmert encoding. The advantage of this test is that we can include a third gender, say "other" gathering PNTS and Other samples. After fitting the model, proceeding to t-tests on each coefficient could show that male and "other" have significantly different likelihoods of sharing disinformation with males being more likely ($t = 7.772$, p-value: $1.106 \cdot 10^{-14}$). It also shows that female does not have significant different means of likelihood of sharing disinformation than male and "other" gathered ($t = -1.611$, p-value: 0.107). Hence, hypothesis testing can sometimes give surprising and contradictory results. Therefore, we could try to go further and take a glance at distributions instead of simple means.



(a) Density, likelihood sharing



(b) Density, likelihood truth

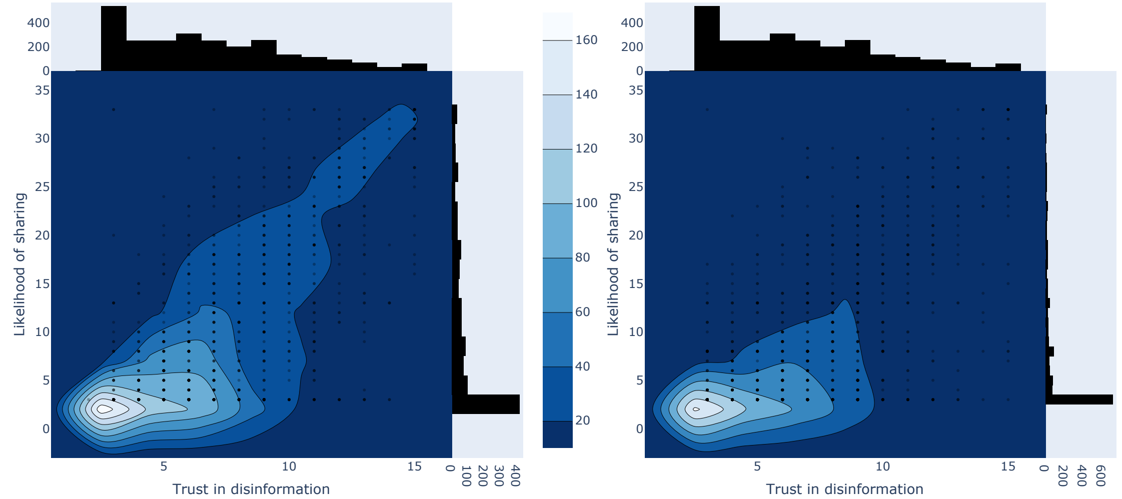
Figure 3.1: Comparison plots between conditional distributions of likelihood sharing and likelihood truth given gender.

Figure 3.1 shows that conditional distributions for likelihood of sharing disinformation and likelihood of trusting false information messages given both genders are relatively different. Indeed, even if the overall shape of the densities are the same, we note that men have an excess of likelihood of sharing disinformation online for high likelihoods. This could be interpreted by many ways, but two main reasons could be either that men have higher trust in disinformation content online or that they share more easily content on social networks than women. However looking at the density plot for the likelihood of trust, it could be inferred that men are more likely to trust disinformation content but it is not as clear as for likelihood of sharing. We may then conduct two two-sample Kolmogorov-Smirnov tests, one for each likelihood in order to quantify these differences. For the likelihood of sharing, the KS test statistic is $D = 0.146$ which is associated with a $3.330 \cdot 10^{-12}$ p-value: therefore, distributions of likelihood of sharing fake news online for both genders are significantly different. For the likelihood of trusting disinformation content, the KS test statistic is $D = 0.085$ which is associated with a $1.846 \cdot 10^{-4}$ p-value: thus, one could assess the same conclusion for the distributions of trusting disinformation content between genders although it is not as clear-cut since the p-value is higher.

Although the data does not contain any information about the precise topic of each message shown to participants, it is specified that messages are politically oriented. And so far, the conclusions we draw agree with [2] in which the author precises that men are more likely to share and trust disinformation content about politics than women. This can be clearly shown on Figure 3.2: men seem to be more likely to share disinformation contents and they are more likely to trust in them, whereas women share less disinformation content and they believe less in them. Correlation must not be interpreted as causality but it could represent one reason (among others) to explain differences in behaviours between both genders on sharing disinformation on social networks. Overall, this plot shows that the majority of participants whatever their gender are not likely to share disinformation online. Therefore, gender could be an important variable to model the likelihood of sharing disinformation on social networks and to better understand this behaviour. However, modelling it only using gender may not be sufficient and others variables should be used. Then, importance of gender in the context of other variables may vary.

3.2 Modelling the likelihood of sharing disinformation

As mentioned previously, overall, most people rated their likelihood of sharing disinformation content as relatively low as shown on Figure 2.3. However, using all available features to model the likelihood of sharing disinformation on social networks seems consistent. One may then fit a normal linear model to the likelihood of sharing disinformation using all features together in order to better understand how these covariates can help in modelling this phenomenon. In order to check for any overfitting as there are numerous features, these estimates were obtained by fitting the model to a train set of 80% of total data. The test set will be used for comparing the error and assess goodness of fit. The reported estimates are shown in Table 1. Note that the reported p-values are derived from t-tests where the null hypothesis is



(a) Male (b) Female

Figure 3.2: Empirical joint density of likelihood of sharing disinformation on social networks and likelihood of trusting disinformation content for both genders.

that the associated coefficient is equal to zero.

The linear model fit procedure gives a R^2 score of 0.574 on the train set and 0.535 on the test set, as well as a mean squared error of 0.0314 on train set and 0.0300 on test set, which are similar values and therefore assess goodness of fit and no apparent overfitting. Removing redundant features helps avoid singularities in the covariance matrix. Results displayed in Table 1 shows that both message and person characteristics are important to model the likelihood of sharing disinformation online. First considering message characteristics: important features according to their significance scores are whether the participant has shared disinformation (with or without knowing), his/her trust in the message and whether he/she had seen this information before. Moreover, Facebook seems to be more impacted by disinformation than Twitter, itself more than Instagram. Regarding person characteristics, males are more likely to share disinformation online, participants from the US have shown to be more likely to share disinformation online than the ones from the UK and critical prosumption increases with the likelihood of sharing fake news online. Some other features help improving the model but are less significant such as extraversion, social conservatism, functional consumption, education, age and employed.

Surprisingly, some features one could think as relevant turn out not to be as significant as expected, such as message source and popularity. Moreover, political beliefs do not seem to have much importance neither. The most important features are trust and whether participants had seen such information before: this shows that people often share content in which they believe to enforce their pre-existing beliefs.

To study any bias and in particular to check if our model is efficient on a gender but not the other, one may have a look at residuals for each group. Figure 3.3 shows prediction points out of sample (test set) according to their predicted values. The global fit seems good, although a lot of data points are stacked on the bottom of the plot due to majority of people being not likely to share any disinformation. One can see on this plot that there does not seem to be any bias in our model between genders since residuals for one gender are not straightforwardly different from the other. In fact, on test set, the mean squared error for women is 0.0278 and the one for men is 0.0329, and the R^2 score for women is 0.539 and 0.515 for men. Therefore, there is no significant bias in our model between genders and it can deal with both of them equally.

To compare these results to another technique, a common method is to use LASSO regression that is a linear regression model along with a L_1 regularisation term. The coefficient of penalisation term can be tuned using cross-validation, where a k-fold cross-validation with $k = 3$ was used to minimise the mean squared error. Using this optimal value, we then fit a new LASSO regression model to the data. LASSO

	Estimate	Effect	p-value	Significance
(Intercept term)	0.2020	+	0.0	***
Education	-0.0080	−	0.0705	.
Age	-0.0085	−	0.0852	.
Employed	0.0081	+	0.055	.
Frequency of use	0.0027	+	0.526	
Shared (found later)	0.0283	+	0.0	***
Shared (while knowing)	0.0224	+	0.0	***
Likelihood truth	0.1190	+	0.0	***
Likelihood seen before	0.0367	+	0.0	***
Authoritative	-0.0032	−	0.4243	
Consensus	-0.0022	−	0.5824	
Openness	-0.0052	−	0.3047	
Neuroticism	-0.0025	−	0.6278	
Extraversion	0.0112	+	0.0136	*
Conscientiousness	0.0054	+	0.2795	
Agreeableness	-0.0076	−	0.1087	
Social	0.0116	+	0.0378	*
Economic	0.0030	+	0.5741	
Functional Consumption	-0.0132	−	0.0322	*
Critical Consumption	0.0101	+	0.1074	
Functional Prosumption	-0.0055	−	0.3819	
Critical Prosumption	0.0168	+	0.005	**
Female	-0.0151	−	0.0005	***
UK	-0.0491	−	0.0	***
Left	0.0066	+	0.1843	
Right	0.0066	+	0.1695	
Facebook	0.0296	+	0.0	***
Twitter	0.0157	+	0.0017	**

Table 1: Estimates of linear model coefficients with likelihood of sharing disinformation as the response. Significance levels correspond to each p-value: ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, .: $p < 0.1$.

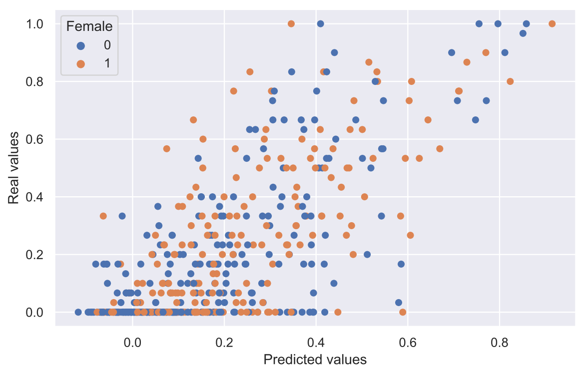


Figure 3.3: Prediction on test set and their actual values.

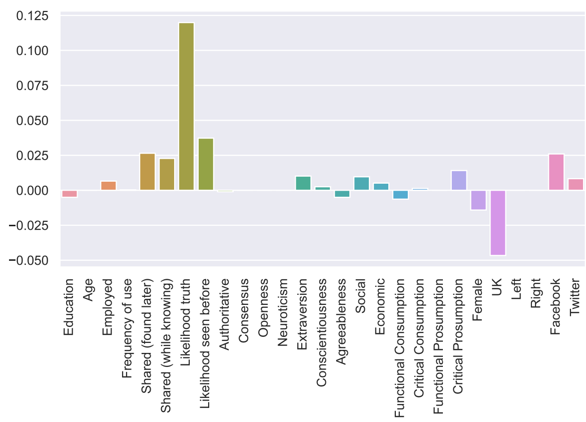


Figure 3.4: Parameters estimates using a LASSO regression model.

penalisation term tends to set *non important* parameters estimates to 0 and thus does features selection. The obtained estimates are shown on Figure 3.4 which gives a visualisation for each feature importance. Again, truth in the disinformation message seems to the most important feature followed by the location (UK or US). Education has a negative component which means that more educated people may be less likely than less educated ones to share disinformation on social networks.

Conclusion

The likelihood of sharing disinformation on social networks depends on many features. Some variables that might be thought to be important, such as an individual's political orientation (except the conservative side) or the consensus around a message and the source reliability actually have little influence. In fact, a determinant of the sharing of fake news on social networks is the trust that people have in the messages they relay in order to increase their visibility and confirm their pre-existing thoughts. Sharing disinformation affects all ages, the level of education plays a slight role and Facebook is the most affected platform. It also showed that men are more affected than women.

This study has some limitations. First, data consist of self-reported likelihoods of sharing disinformation on social networks and therefore does not reflect exactly participants real-world behaviour. Hence, it is assumed that what people self-report is similar to what they would do online. Also, data shows that only a few people share disinformation online since the distribution of likelihood of sharing is highly right-skewed. Thus, among all collected samples, around the half had perfectly identified disinformation messages which leaves few examples of people who would be unlikely to share, and even fewer of those who would actually share. Finally, one of the main issue with this work is that the most important variable for modelling the likelihood of sharing is the likelihood of trust which is in practice very hard to measure. In real-world applications, social networks platforms could not evaluate people trust in a message. Removing this feature from the models decreases the R^2 score to 0.37 which is still a correct model but not as efficient as previously.

Considered models here are relatively simple but therefore explainable. On the one hand, one could have used more complex models such as ensemble learning methods. These models could then better understand the structure of the data like non linearities and give better results in terms of mean squared error for example but they would be less explainable. On the other hand, another way of improvement could be to see this problem as a binary classification problem, where the two classes correspond to either people that are likely or not to share disinformation on social networks and where the split is done using a threshold. A such classifier has been fitted to the data trying to classify people with a likelihood of sharing disinformation online greater than 0.15 (40% of the data) from the others, using a boosted trees classifier which parameters were tuned using cross-validation on a grid and obtained an accuracy of almost 80% on the test set, a recall of 81% and a precision of 90%.

References

[1] Chengcheng Shao et al. (2017) *The spread of misinformation by social bots*, URL

[2] Ester Almenar et al. (2020), *Gender Differences in Tackling Fake News: Different Degrees of Concern, but Same Problems*, Media and Communication, URL

[3] Tom Buchanan (2020) *Predictors of likelihood of sharing disinformation on social media*, URL

[4] Chris Marsden et al. (2021) *Responding to Disinformation, Ten Recommendations for Regulatory Action and Forbearance*, URL

[5] Tom Buchanan (2020) *Why do people share disinformation on social media?*, School of Social Sciences, University of Westminster, London, UK URL

[6] Andrew Guess et al. (2019) *Less than you think: Prevalence and predictors of fake news dissemination on Facebook*, Science Advances URL

[7] Tom Buchanan (2021) *Trust, personality, and belief as determinants of the organic reach of political disinformation on social media*, School of Social Sciences, University of Westminster, London, UK URL