What Leads People to Make Health-Related Decisions on Social Media

I. Abstract

In current situation of vigorous spread of unverified information, also more interacting on social media, many number of people would make decisions based on health information from the social media. We set three models, dependent variable fixed, more independent variables regrading usage of social media and our main variable are added through model 2 and model 3. We observed our model 2 and 3 illustrated improved performance in terms of R squared and less MSE. The more people engage on social media, the more they refer health information on social media. If the respondents use health information in discussions with their health care provider, they were observed highest tendency to use health information from social media. The results gave the possibility that regardless of personal perceptions, more exposure to health information on social media can directly induce users to apply the information without considering the reliability.

II. Introduction

In the era of easy-to-get information floating around, lots of unverified information is being shared from people to people via social media. We had focused on the possible factors that can affect people on decision making regarding health information, having more closer look to usage of social media. In this situation, lively engagement on social media has increased, leaving the increased exposure to the wrong information. This paper reports the relationship between false or misleading health information can affect people to make decisions from the social media. This study highlights the impact on decision making regarding health information especially from the social media.

2.1 Purpose

The purpose of this case study is to observe how exposure to health misinformation on social media affects health related decision making. Unlike previous studies that focused on perception-based surveys, this research takes a quantitative approach to measure the impact of health information on social media.

2.2 Background

Background.

2.3 Significance of this Study

To warn people not to absorb health information without any barriers. To explore what factors affect people related to this issue, we can prevent false or misleading information from being influence on decision making which can be fatal or critical.

III. Methodology

3.1 Data Source

The dataset is from HINTS(Health Information National Trend Survey), 2022 survey. Total responds are 6,252 and complete responds are 6,185 and other 67 responds are partially responded. We converted the given R Data(.rda) into CSV file(.csv) for handling with programming on python.

3.2 Programming Environment

- Environment: WSL2, Ubuntu 20.04
- Programming Language: Python, version 3.11.8
- TensorFlow Version: 2.12.0

3.3 Variable Descriptions

Table 1

Table 1: Variable Descriptions

Variable	Survey Question	Scale
‘SocMed_MakeDecisions’ (Dependent Variable)	B14 a. I use information from social media to make decisions about my health	Ordinal
‘MisleadingHealthInfo’ (Main Independent Variable)	B13. How much of the health information that you see on social media do you think is false or misleading?	Ordinal
‘Age’ (combined as PC)	R1. What is your age	Ratio
‘IncomeFeelings’ (combined as PC)	R15. Which one of these comes closest to your own feelings about your household’s income?	Ordinal
'Age_Income_PC1'	Combined variable with ‘Age’ and ‘IncomeFeelings’ through PCA(Principal Component Analysis)	Ordinal
‘MaritalStatus’	R6. What is your marital status?	Nominal (Binary)
‘BirthGender’	R2. On your original birth certificate, were you listed as male or female?	Nominal (Binary)

‘SocMed_DiscussHCP’	B14 b. I use information from social media in discussions with my health care provider	Ordinal
‘SocMed_TrueFalse’	B14 c. I find it hard to tell whether health information on social media is true or false	Ordinal
‘SocMed_SameViews’	B14 d. Most of the people in my social media networks have the same views about health as me	Ordinal
‘SocMed_SharedPers’	B12 b. Shared personal health information on social media	Ordinal
‘SocMed_SharedGen’	B12 c. Shared general health-related information on social media (for example, a news article)	Ordinal
‘SocMed_WatchedVid’	B12 e. Watched a health-related video on a social media site (for example, YouTube)	Ordinal

Note. SocMed stands for Social Media, HCP stands for Health Care Provider.

Table 2

Table 2: Original & Recorded Response Counts

Variable	Original Response	Recorded Response	Original Counts	Recorded Counts
MisleadingHealthInfo	I do not use social media	-1	1211	0
	A little	0	855	518
	Some	1	2256	1484
	A lot	2	1740	1169
Age	For each age	Only Type Casted	6154	4248
IncomeFeelings	Finding it very difficult on present income	0	346	148
	Finding it difficult on present income	1	763	367
	Getting by on present income	2	2140	1127

	Living comfortably on present income	3	2518	1529
Age_Income_PC1*		-3		12
		-2		143
		-1		466
		0		2049
		1		492
		2		9
BirthGender	Male	0	2307	1297
	Female	1	3535	1874
MaritalStatus	Single, never been married	0	1119	1614
	Separated	0	136	
	Widowed	0	646	
	Divorced	0	939	
	Living as married or living with a romantic partner	0	373	
	Married	1	2624	1557
SocMed_DiscussHCP	Strongly disagree	0	2947	2534
	Somewhat disagree	1	977	851
	Somewhat agree	2	903	776
	Strongly agree	3	108	87
SocMed_TrueFalse	Strongly disagree	0	818	690
	Somewhat disagree	1	885	762
	Somewhat agree	2	1784	1543
	Strongly agree	3	1441	1253
SocMed_SameViews	Strongly disagree	0	1137	267
	Somewhat disagree	1	1552	1377
	Somewhat agree	2	1847	1652
	Strongly agree	3	310	267
SocMed_SharedPers	Never	0	5161	3485
	Less than once a month	1	604	516
	A few times a month	2	171	141
	At least once a week	3	73	54
	Almost every day	4	58	52
SocMed_SharedGen	Never	0	4305	2868
	Less than once a month	1	1238	1064
	A few times a month	2	402	340
	At least once a week	3	134	110
	Almost every day	4	59	48
SocMed_WatchedVid	Never	0	2685	1390
	Less than once a month	1	1836	1491
	A few times a month	2	1047	876
	At least once a week	3	419	350
	Almost every day	4	171	141

Note. Invalid data such as missing data, incomplete data, multiple responses selected data and data with technical issues was excluded from counting for both before and after preprocessing.

*Note. Value of combined variable can be positive(negative) if both age and the satisfaction on current income have relatively higher(lower) value than average or either one has very large(small) value. Both values are around the average if the value is zero.

Table 3

Table 3: Correleation Matrix (Model 3)

	PC	BG	MS	Dis_HCP	TF	SV	ShrPer	ShrGen	VID	MLHI
PC	1.000	-0.073	0.176	-0.038	0.043	-0.022	-0.105	-0.076	-0.163	0.029
BG		1.000	-0.127	0.025	-0.004	0.014	0.020	0.033	0.017	-0.001
MS			1.000	-0.016	0.012	0.023	-0.029	-0.003	-0.007	0.022
Dis_HCP				1.000	0.045	0.153	0.230	0.277	0.296	-0.201
TF					1.000	0.137	-0.016	-0.041	-0.031	0.080
SV						1.000	0.089	0.126	0.078	-0.043
ShrPer							1.000	0.503	0.271	-0.117
ShrGen								1.000	0.385	-0.112

VID									1.000	-0.155
MLHI										1.000

Note. PC= Age_Income_PC1, MS=MaritalStatus, BG=BirthGender, Dis_HCP= SocMed_DiscussHCP, TF= SocMed_TrueFalse, SV= SocMed_SameViews, ShrPer=SocMed_SharedPers, ShrGen=SocMed_SharedGen, VID=SocMed_WatchedVid, MLHI=MisleadingHealthInfo

Two variables 'SocMed_SharedPers' and 'SocMed_SharedGen' have high correlations which means shared personal information and general information respectively. On the other hand, variable 'Dis_HCP', 'SocMed_DiscussHCP' has relatively high correlations with responses regarding social media. Main independent variable of this research, 'MisleadingHealthInfo' showed low correlations between all the other variables.

3.4 Preprocessing

All the variables are numerated for proceeding regression. Responses are converted in numbers zero to (number of valid answers – 1) for each, except the variable named 'MisleadingHealthInfo'. In the middle of converting process. To exclude responses who answered 'I do not use social media', this response is numerated as '-1'. Response counts are all checked using 'value_counts()' function to validate the process.

Dependent variable for this research consists of answers from whom not responded "I do not use social media" from survey question B13, variable name 'MisleadingHealthInfo'. So the response 'I do not use social media' was converted into negative number(-1), and changed into NaN(Not a Number) to drop invalid value at once in later step. To secure the completeness of data after dropping the responses, exclusion of this response proceeded at first. In the same manner, after checking all the valid responses are turned into integers, we dropped the rows with NaN values and all the string values which includes missing data and other invalid data. The codes for typecasting to integer data type to make sure the regression step to be conducted with no errors.

Since the two variables, Age and IncomeFeelings(satisfaction on current income), exhibit high multicollinearity, they were combined into a single component using Principal Component Analysis (PCA). First, the variables were standardized using a StandardScaler, and then PCA was applied to extract one principal component, which is now represented as Age_Income_PC1.

3.5 Analysis Methodology

We coded with Python programming language, and converted R Data file into CSV file as noted. Data frame from Pandas stored the CSV file data as a data frame data type. Structure for this research is comparing three models, in terms of performance in R-squared measurements. To see the results at once, iterations are used and the results for each model are saved and print the results after the iteration step.

To dependent variable is fixed, independent variables are added for each step. Model 1 has 6 independent variables(one is combined with two variables) and model 2 has three more independent variables from the same survey question B12, all sub-questions are asking interaction with social media. In the middle, the variable named 'SocMed_Visited' and 'SocMed_Interacted' were excluded due to the high multicollinearity between other variables. Hence Model 2 has 9 independent variables. Model 3 has one more independent variable, which is our main variable to see the effect on the dependent variable. For each iteration, linear regression model is created and learned for independent variables of model 1, model 2, and model 3 respectively. Performance was measured as R-square score and MSE(Mean Squared Error). To see the coefficients and other factors, the linear regression model was fitted using OLS.

Checking VIFs and cross validation process were to check reliability of the models. For the deeper assessment on this model, results from ordinal logistic model and TensorFlow was recorded too. The accuracy of TensorFlow Model was referenced, and visualized plot from the result of TensorFlow Model for overfitting was also utilized as well as residual histogram of linear regression.

IV. Results

Table 5

Table 4: Model Performance Results

Models	Model 1	Model 2	Model 3
R-Squared	0.328571	0.362021	0.381289
	(-)	(10.2%)	(5.32%)
MSE	0.441093	0.419118	0.406460
	(-)	(4.98%)	(3.02%)
Pseudo	0.186539	0.205195	0.220072
R-squared	(-)	(10.0%)	(7.25%)
TensorFlow	0.667059	0.670588	0.684706
Accuracy	(-)	(0.52%)	(2.11%)

Test results showed improvement in terms of performance on R-squared score and a reduction in errors(MSE) as step goes further. But the low absolute value of the results have potential improvements.

R-squared is also known as Coefficient of Determination, an indicator that shows how the model explains the volatility of dependent variables. The value lies between 0 and 1, the model explains well if the value is closer to 1.

MSE, Mean Squared Error quantifies the prediction failures by getting the results from the difference between the actual and predicted value. The difference is squared to prevent cancellation of errors in the positive and negative directions, and then averaged. A lower MSE indicates better prediction accuracy by the model.

Pseudo R-squared is conceptually similar to the R-squared in linear regression, it does not represent the proportion of variance explained by the model. Instead, it evaluates the improvement of the model compared to a baseline (null) model. The value of Pseudo R-squared lies between 0 and 1. A higher value indicates that the model is better at explaining the relationship between the independent and dependent variables. But we could observe tendency of improvement at prediction as the model gets more related independent variables.

Accuracy of TensorFlow model is the portion of number of correct predictions divided by total number of predictions. This model has more than 66% probability of prediction and improvement is also observable.

4.1 OLS Regression Results

Table 6

Table 5: Linear Regression Results(Model 1)

Variables	coefficients	standard errors	t-value	P>|t|
const	0.1592***	0.031	5.116	0.000
Combined component with age and satisfaction on current income	-0.0676***	0.015	-4.411	0.000
Now married or not	-0.0146	0.022	-0.652	0.515
Birth gender	-0.0178	0.022	-0.806	0.421
Use health information on social media on discussion with health care provider	0.5038***	0.013	38.918	0.000
Hard to tell whether health information on social media is true or false	-0.0037	0.011	-0.349	0.727
Thinks most people in my social media have the same views about health as me	0.0618***	0.013	4.938	0.000

Note. *p < .05 *, p < .01**, p <.001 ****

Variables with statistical significance(denoted with stars) are interpreted that the variables have significant influence on model's prediction. Equation of the linear regression can be written as follows:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

Beta(β) is coefficient for each variable, while β_0 is constant and ϵ is error term. Dependent variable, y, which responses zero to three(strongly disagree, somewhat disagree, somewhat agree, and strongly agree) can be predicted as we put numerated independent variables multiplied by each coefficient by the linear regression model. Combined component, how the people think others watch health related information on social media, and the constant term helps model to predict more precisely. Importantly, the more people use health information from the social media to discuss with their health care provider, they make decisions about their own health based on health information on social media, by the result.

Table 7

Table 6: Linear Regression Results(Model 2)

Variables	coefficients	standard errors	t-value	P>|t|
const	0.0607	0.032	1.901	0.057
Combined component with age and satisfaction on current income	-0.0359**	0.015	-2.341	0.019
Now married or not	-0.0160	0.022	-0.727	0.467
Birth gender	-0.0276	0.022	-1.265	0.206
Use health information on social media on discussion with health care provider	0.4559***	0.013	33.776	0.000
Hard to tell whether health information on social media is true or false	0.0004	0.010	0.042	0.966
Thinks most people in my social media have the same views about health as me	0.0558***	0.012	4.516	0.000
Shared personal health information on social media	0.0478**	0.018	2.658	0.008
Shared general health-related information on social media	0.0181	0.016	1.163	0.245

Watched a health-related video on social media site	0.1002***	0.011	9.033	0.000

Note. *p < .05 *, p < .01**, p <.001 ****

Model 2 has more independent variables regarding usage of social media. In addition to the observation from the model 1, the more people share their information, especially personal one than general, and watch health related video on social media, they use health information on social media to make decisions in terms of health.

Table 8

Table 7: Linear Regression Results(Model 3)

Variables	coefficients	standard errors	t-value	P>|t|
const	0.2537***	0.037	6.881	0.000
Combined component with age and satisfaction on current income	-0.0357*	0.015	-2.363	0.018
Now married or not	-0.0157	0.022	-0.725	0.468
Birth gender	-0.0241	0.021	-1.124	0.261
Use health information on social media on discussion with health care provider	0.4351***	0.013	32.326	0.000
Hard to tell whether health information on social media is true or false	0.0088	0.010	0.856	0.392

Thinks most people in my social media have the same views about health as me	0.0536***	0.012	4.399	0.000
Shared personal health information on social media	0.0389*	0.018	2.194	0.028
Shared general health-related information on social media	0.0205	0.015	1.334	0.182
Watched a health-related video on social media site	0.0902***	0.011	8.224	0.000
How much of health information on social media is false or misleading	-0.1538***	0.015	-10.068	0.000

Note. *p < .05 *, p < .01**, p <.001 ****

Model 3 has one more variable compared to Model 2, our main variable that how the people think health information on social media is false or misleading. We can highlight that it has negative relationship on decision making and negative perception when it comes to health. And the absolute value is secondly significant excluding the constant term.

Table 9

Table 8: Linear Regression Coefficients for Models 1, 2 and 3

Variables	Model 1	Model 2	Model 3
const	0.1592***	0.0607	0.2537***
Combined component with age and satisfaction on current income	-0.0676***	-0.0359**	-0.0357*
Now married or not	-0.0146	-0.0160	-0.0157

Birth gender	-0.0178	-0.0276	-0.0241
Use health information on social media on discussion with health care provider	0.5038***	0.4559***	0.4351***
Hard to tell whether health information on social media is true or false	-0.0037	0.0004	0.0088
Thinks most people in my social media have the same views about health as me	0.0618***	0.0558***	0.0536***
Shared personal health information on social media		0.0478**	0.0389*
Shared general health-related information on social media		0.0181	0.0205
Watched a health-related video on social media site		0.1002***	0.0902***
How much of health information on social media is false or misleading			-0.1538***

Note. p < .05 *, p < .01**, p <.001 ***

This table shows how each variable’s coefficient is calculated by each model. The significance of the combined factor and the variable about how they shared personal health regarded information on social media are diluted as we put more independent variables but still it has non-negligible importance. Using health information from the social media on discussion with health care providers, thinking most people in their social media have same views about health as them, watching health-related video on social media, and how much people think health information on social media is false or misleading have strong relationship with how they can affect people to make health related decisions with the health information on social media.

4.2 Ordinal Logistic Regression Results

Table 10

Table 9: Ordinal Logistic Regression Coefficients for Models 1, 2 and 3

Variables	Model 1	Model 2	Model 3
const	-0.1904	-0.0947	-0.1052
Combined component with age and satisfaction on current income	-0.0447	-0.0469	-0.0468
Now married or not	-0.0569	-0.0738	-0.0626
Birth gender	1.4757	1.3431	1.2966
Use health information on social media on discussion with health care provider	-0.0216	0.0047	0.0485
Hard to tell whether health information on social media is true or false	0.2779	0.2507	0.2515
Thinks most people in my social media have the same views about health as me		0.1317	0.1012
Shared personal health information on social media		0.0742	0.0832
Shared general health-related information on social media		0.3492	0.3304
Watched a health-related video on social media site			-0.6161
0/1	1.8107	2.2115	1.5466
1/2	0.4577	0.4862	0.5142
2/3	1.2376	1.2671	1.2833

Note. *p < .05 *, p < .01**, p <.001 ****

Ordinal logistic regression tells us the probability what category the response will be placed. The concept of statistical significance is applied as the same way. By using this model, we could reconfirm the possible effects of variables. Tendency from the linear regression model is not that different with this ordinal logistic model.

The coefficients of 0/1, 1/2, and 2/3 indicate the frontiers (or thresholds) that separate the response categories of the y-variable in ordinal logistic regression. These thresholds are the boundaries at which the cumulative probabilities of being in a specific category transition from one category to the next.

4.3 Validation and Reliability

Table 11

Table 10: VIF Results

Variables	Model 1	Model 2	Model 3
Combined component with age and satisfaction on current income	1.057335	1.104089	1.104516
Now married or not	2.090433	2.142158	2.271786
Birth gender	1.712275	1.755736	2.271786
Use health information on social media on discussion with health care provider	1.541809	1.772711	1.792083
Hard to tell whether health information on social media is true or false	2.913990	2.981806	3.432874
Thinks most people in my social media have the same views about health as me	2.792179	2.880870	2.974318
Shared personal health information on social media		1.537559	1.541718
Shared general health-related information on social media		2.050138	2.052914
Watched a health-related video on social media site		2.517968	2.527232
Combined component with age and satisfaction on current income			2.923943

VIF, Variance Inflation Factor, indicates multicollinearity for each independent variable. Correlated variables have high VIF value which can distort the result. Two variables, age and satisfaction on current income are combined into one variable as PC(Principal Component) since they had value over 7, indicating it may have multicollinearity.

Table 12

Table 11: Cross Validation Check Results

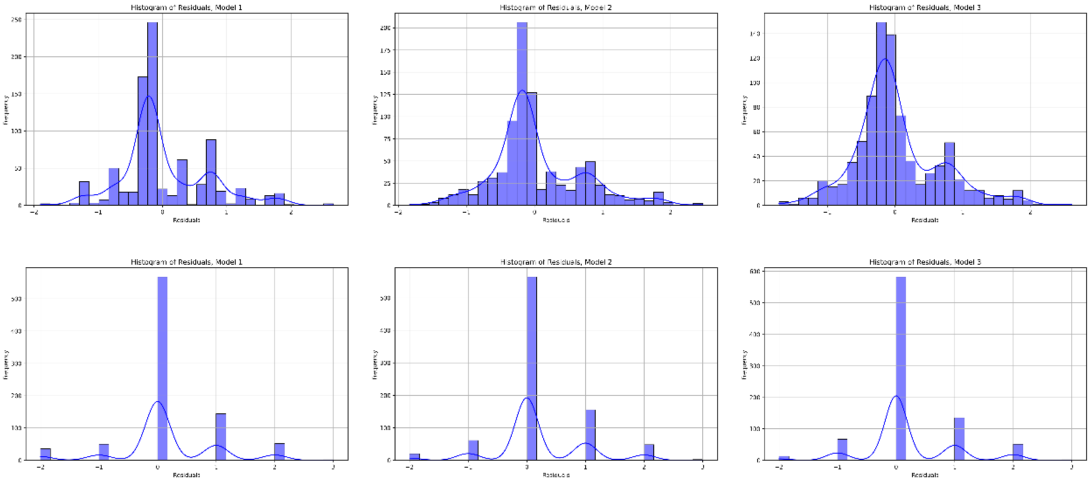
	Model 1	Model 2	Model 3
Standard Deviation of CV scores	0.025831	0.022424	0.021184

Reliability was also check through cross validation test. The test were conducted five times for each model and reported quite low difference within a trial which can guarantee reliability of this model in some degree. CV scores are from MSE(Mean Squared Error) of the linear regression model.

Residual Histogram

Figure 1

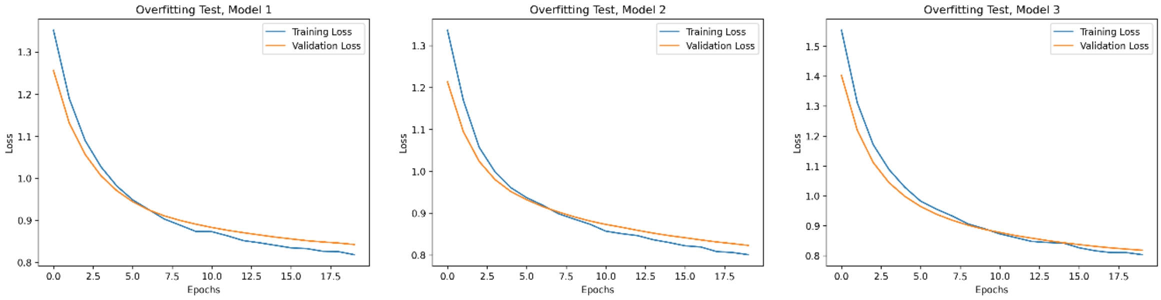
Figure 1: Residual Histograms



This plot shows residuals in our regression model. Histogram on first row is from linear regression model and second row is from ordinal logistic model. Histogram illustrates distribution of differences between the observed values and predicted values. The residuals are not skewed and seem to follow normal distribution to have bell-shaped curve. To visualize probability of the result from the ordinal logistic model, the results are selected only on response with the highest probability, having discrete value. Large portion of values are plotted in the middle as illustrated, it means that predictions have high accuracy since residual stands for the difference between predicted and actual category. But if it is skewed to zero, it could be a signal for overfitting.

Figure 2

Figure 2: Overfitting Test Results



This figure illustrates whether the model is overfitted. The results are derived from a TensorFlow model. After the regularization, the curve became smoother and the difference between the graphs are quite little. With the fact that both graphs illustrate decreasing losses and two graphs are almost aligned, this model has less possibilities of overfitting problem.

V. Discussion

Through this research, we observed our main variable, how people think health information on social media is false or misleading has negative correlation with health-related decision making based on information from social media. By this result, we can consider people with negative perception on health information spreading on social media are less likely to accept health information without any verification.

As we see the improvement from model 1 to model 2 is greater than model 2 to model 3, more interaction on social media can affect more on health related decision making than its perception.

VI. Conclusion

We checked possible danger that more engagement on social media can make people unconsciously follow any information floating around social media. This research had a closer look at health information, which can be fatal and should have consideration on accepting it.

VII. Appendix

Appendix A

Data Set Used for This Research

Data set used for this research can be downloaded here. Sign-in is needed. You can use your email address to enter the download page. Or, you can directly download the zip file by clicking here. The data set is R data and supporting documents (ZIP, 16.7 MB) from HINTS 6 (2022) dataset, updated May 2024.

Appendix B

Python Codes

The programming code for this research is uploaded here. Modeling was done with the file ‘main.py’. Response counts were executed using the file ‘response_counts.py’ to store the counts in CSV file.

Appendix C

Tables and Figures