ANALYSIS OF ONLINE NEWS POPULARITY DATA IN MACHINE LEARNING REPOSITORY USING R STUDIO

JOHN MICHAEL D. AQUINO
https://orcid.org/0000-0002-0253-1935
johnmichael.aquino@lspu.edu.ph
Laguna State Polytechnic University
Sta. Cruz, Laguna, Philippines

ABSTRACT

Media are tools to get information across the globe. And online news is described as an article that is made for spreading awareness on all kinds of topics or issues published on the internet. In this research, the researcher wanted to analyze the online news popularity data in a machine learning repository using the R Studio in terms of the relationship of variables to identify the popularity of online news articles. Specifically, the researcher desired to identify the results in Principal Component Analysis using the Online News Popularity data in Machine Learning Repository; *show the results in the Path Analysis using the online news popularity mined data; reveal the relationship of the variables (publication article, words, and links) towards the popularity of online news; display the structural equation model derived from the mined data; and determine the relationship between confirmatory factor analysis and the structural equation model. The data consisted of a heterogeneous set of features which was extracted into 61 characteristics (58 predictive features, two non-predictive, and one goal field) and from the overall of 39,644 respondents. The mined data set came from UCI Machine Learning Repository in the CSV file. This study employed a quantitative method. RStudio is software used which was a computational research tool for creating dynamic and reproducible research. R software is installed in the computer and used to treat and analyze the data in all statistical requirements of the study.*

Keywords: Principal Component Analysis, Structural Equation Model, Path Analysis, Confirmatory Factor Analysis, Online News Popularity

INTRODUCTION

Media are tools to get information across the globe. This is an avenue to store and deliver information or data in different aspects like news media, print media, advertising, digital media, and others. In addition, online news is described as component of media where citizens can find an authentic information (Wicaksono and Supianto, 2018). People nowadays are using gadgets to get information worldwide with the help of the internet. Using various applications like Facebook, Twitter, and Instagram, people can be updated on what happens in the community. Rathord, Jain and Agrawal (2019) stated that online news can easily spread across the globe. And people are habitually sharing and reading online news in media. Likewise, online news is described as an article which made for spreading awareness on all kinds of topics or issues published on the internet (Choudhary, Sandhu, and Pradhan, 2017). With this, it tends to be popular if the published article was interesting to readers.

The fame of online news articles identified founded on the quantity of shares (Ren and Yang, 2015). Today, because of this pandemic, the center of people's entertainment focuses on them accessing their gadgets to read and share online news. Online news popularity as a research topic becomes the new trend due to Web expansion (Fernandes, Vinagre, and Cortez, 2015). The

researchers perceived that this topic is more interesting to the readers. With this, Guan, Peng, Li, and Zhu (2017) attested that online news has a huge growth of information as well as it strengthens and increases the attention of the readers in today's period. R studio is described as statistical tools used for analysis of data and reproducible research (Loraine, et al., 2015). This helps the world to build an open-source software for various kinds of researches. The primary purpose of this software is to threat large data to analyze easily and there are multiple ways to interface this R. In addition, various fields of study as well as to different practitioners could use this software (Kronthaler & Zöllner, 2021).

Moreover, the purpose of this research was to determine the results of Principal Component Analysis (PCA), Path Analysis, and creating Structural Equation Model (SEM) using R studio, and if there is a relationship between the confirmatory factor analysis and structural equation model from large collected data set of UCI Machine Learning Repository with over 39,644 articles.

OBJECTIVES OF THE STUDY

This study aimed to analyze the data to determine the results of Principal Component Analysis (PCA), Path Analysis, and creating Structural Equation Model. Specifically, the researcher wanted to:

1. identify the results in Principal Component Analysis using the Online News Popularity data in Machine Learning Repository;
2. show the results in the Path Analysis using the online news popularity mined data.
3. Reveal the relationship of the variables (publication article, words, and links) towards the popularity of online news;
4. display the structural equation model derived from the mined data; and
5. determine the relationship between confirmatory factor analysis and the structural equation model.

METHODOLOGY

This study employed a quantitative method. R Studio is software used which was a computational research tool for creating dynamic and reproducible research (Gandrud, 2013). This software is suitable for a large data set and all researches in any quantitative discipline.

The data set was from the Mashable website accumulated in two years. This was summarizing as a heterogeneous set of features which was extracted into 61 characteristics (58 predictive features, two non-predictive, and one goal field) and from the overall of 39,644 respondents. In addition, this data set did not share the original content. However, the original content may be seen in the provided URLs.

Table 1
Summary of the Variables

Codes	Characteristics	Features
url and timedelta	2 non-predictive	URL and publication of the articles
a1-a5	Words	Numbers of words of the title, content, average word length, rate of unique, and non-stop words of content
b1-b2	Links	Numbers of links and number of links to other articles
c1-c2	Media	Number of pictures and videos
d1-d20	Keywords	Number of keywords; worst, best, and average (number of shares); category of the articles
e1-e8	Publication Time	Day of the week or weekend
f1-f21	Natural Language Processing (NLP)	Closeness to LDA topics; Text Polarity/subjectivity; Rate and polarity of both positive and Negative words; and Absolute subjectivity and polarity level
g1	Goal	Number of Shares

Table 1 reveals the summary of all the variables in the mined data. It shows the codes used by the researcher, the characteristics of each group as well as the features or information of the attributes.

The mined data set came from UCI Machine Learning Repository in the CSV file. R software was installed in the computer which was used to treat and analyze the data in all statistical requirements of the study.

Table 2
Summary of Descriptive Statistics

Summary	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
a1-a5	-641.000	8.285	9.405	9.710	11.275	22.343
b1-b2	0.00	1.00	4.00	7.59	11.00	298.00
c1-c2	-91.000	0.000	1.000	3.294	3.000	116.000
d1-d20	-843295.3	-5194.8	-2195.5	-6397.1	-976.6	6.1
e1-e8	-1.00000	0.00000	0.00000	0.03711	0.00000	1.00000
f1-f21	-0.98181	-0.13999	0.02501	0.02854	0.14611	0.92699
g1	1	946	1400	3395	2800	843300

The table reveals the descriptive statistics of the group of codes. The mean of g1 stood out among them all which was pegged at 3395. It talks about the number of shares or goal of the mined data. When a1-a5 was discussed, the number of words in the title, content, average word length, rate of unique, and non-stop words of content were at 9.710 and b1-b2 were at 7.59 focused on the number of links. While c1-c2 acquired the mean of 3.294 which intended to digital media both pictures and videos. Meanwhile, e1-e8 has the mean of 0.03711 results of publication time from Monday to weekend. And the Natural Language Processing (NLP) (f1-f21) has 0.02854 acquired from closeness to LDA topics; text polarity/subjectivity; rate and polarity of both positive and negative words; and absolute subjectivity and polarity level.

However, d1-d20 has the mean of -6397.1 which was based on the number of keywords; worst, best, and average (number of shares), and category of the articles.

RESULTS AND DISCUSSION

The mined data was from the Mashable website in two years. This was summarizing as a heterogeneous set of features which was extracted into 61 attributes and from a total of 39,644 respondents.

1. Principal Component Analysis using the Online News Popularity data in Machine Learning Repository

Table 3
Seven (7) Factors exist as suggested from the result of PCA

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
SS loadings	2.993	2.456	2.431	2.100	2.035	1.547	0.978
Proportion Var	0.086	0.070	0.069	0.060	0.058	0.044	0.028
Cumulative Var	0.086	0.156	0.225	0.285	0.343	0.387	0.415

In the Principal Component Analysis results, it suggests seven (factors) that exist. The table presents the data in each factor in three indicators; the sum of squared loadings, proportion variance, and cumulative variance. The factors in the sum of squared loadings were 2.993, 2.456, 2.431, 2.100, 2.035, 1.547, and 0.978. In proportion variance, the factors got 0.086, 0.070, 0.069, 0.060, 0.058, 0.044, and 0.028. Finally, the cumulative variances had the results of 0.086, 0.156, 0.225, 0.285, 0.343, 0.387, and 0.415 in all factors. In addition, the seven factors were sufficient in the test of the hypothesis. The chi-square statistic was 400362.8 on 371 degrees of freedom and the p-value was 0.

2. Path Analysis Using the Online News Popularity Mined Data

Path Analysis is the diagrammatic representation of the theoretical model using standardized notation. It presents regression equations specified between measured variables. It also reveals the effects of predictor variables in the criterion. The dependent variables can be direct, indirect, or total. Garson (2013) attests that path analysis is an extension of the regression model which was used to test the fit of the correlation matrix to be compared by the researcher in the two or more models.

Table 4

The Goodness of Fit Testing Results

Index	Cutoff Value	Result	Notes
CFI	≥0.90	0.999	Good
TLI	≥0.90	0.998	Good
RMSEA	<0.08	0.047	Good
SRMR	<0.08	0.015	Moderately Good

The table shows that the model was good. It estimates the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), the Root Mean Square Error Estimation (RMSEA), and the Standardized Root Mean Square Residual (SRMR). The table al so presented the cut-off value in each indicator of the model as well as the result

Table 5

Regressions

Regressions						
	Estimate	**Std. Err**	**z-value**	**P(>|z|)**	**Std.lv**	**Std. all**
timedelta~						
a1	-24.091	NA			-24.091	-0.238
a2	0.000	NA			0.000	0.000
a3	-46.785	NA			-46.785	-0.769
a4	-149.409	NA			-149.409	-3.650
a5	289.896	NA			289.896	4.419
g1~						
b1	49.156	NA			49.156	0.048
b2	-83.844	NA			-83.844	-0.028
c1	43.043	NA			43.043	0.031
c2	64.927	NA			64.927	0.023
d1	-411.075	NA			-411.075	-0.030
d2	96.487	NA			96.487	0.016

Moreover, the results of path analysis using lavaan reveals that in standard deviation as a whole of regressions in Table 5 between time delta and a1 to a5 was very low. However, a5 (rate of unique non-stop words in the content) got the highest which has 4.419 and the lowest was a1 (number of words in the title) which has -0.238. While the regression from g1 to b1-d2, b1 (number of links) stood out with 0.048, and d1 (average length of words in the content) is the lowest which acquired -0.030.

Table 6
Covariance, Variances, and R-square

Covariance:						
	Estimate	**Std.Err**	**z-value**	**P(>|z|)**	**Std. lv**	**Std. all**
timedelta~~g1	1.656	NA			1.656	0.000
Variances:						
	Estimate	**Std.Err**	**z-value**	**P(>|z|)**	**Std. lv**	**Std. all**
.timedelta	42588.808	NA			42588.808	0.929
.g1	134507818.342	NA			134507818.342	0.995
R square						
	Estimate					
timedelta	0.071					
g1	0.005					

Furthermore, the covariance of time delta to g1 (number of shares) has the standard deviation of 0.000, the variances of time delta with 0.029, and g1 (number of shares) with 0.995. Meanwhile, in R square, it is estimated that time delta has 0.071 and g1 (number of shares) has 0.005.

3. Relationship of the Variables towards the Popularity of Online News

The researcher analyzes data in which the accuracy of the results was tested. The researcher also used the confirmatory factor analysis using the R studio to verify the scales used to assess the validity of both convergent and discriminatory. With this, using the mined data, the researcher wanted to identify the relationship of the publication article, words, and links towards the popularity of online news.

Table 7
The Goodness of Fit Testing Results

Index	Cutoff Value	Result	Notes
CFI	≥0.90	0.989	Good
TLI	≥0.90	0.979	Good
RMSEA	<0.08	0.148	Moderately Good
SRMR	<0.08	0.014	Moderately Good

The testing results in Table 7 estimates the Comparative Fit Index (CFI) with the result of 0.989, Tucker-Lewis Index (TLI) which pegged at 0.979, Root Mean Square Error Estimation (RMSEA) acquired 0.148, and Standardized Root Mean Square Residual (SRMR) had 0.014 which tells the goodness of the model.

Table 8
Latent Variables

Latent Variables						
	Estimate	**Std. Err**	**z-value**	**P(>|z|)**	**Std.lv**	**Std. all**
Publication=~ timedelta	1.000				214.161	1.000
Words=~						
a3	1.000				3.520	1.000
a4	1.485	0.000	6800.772	0.000	5.229	1.000
a5	0.927	0.000	11548.767	0.000	3.264	1.000
Links =~						
b1	1.000				5.934	0.524
b2	0.492	0.732	0.672	0.501	2.919	0.757

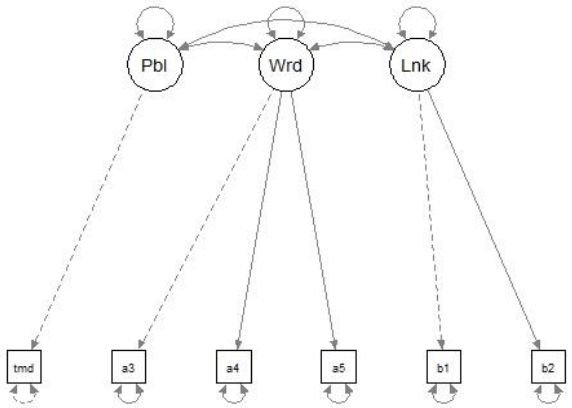
The Std. all columns present the data that is standardized so that both latent variables and observed variables had a variance of one as can be seen in Table 8.

Table 9
Regressions, Covariance, Variance, and R-square

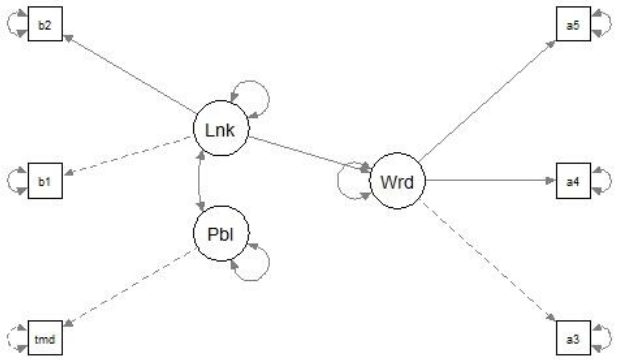
Regressions						
	Estimate	**Std. Err**	**z-value**	**P(>|z|)**	**Std.lv**	**Std. all**
Words~						
Links	0.004	0.004	0.936	0.349	0.006	0.006
Covariance						
Publication~						
Links	0.474	8.043	0.059	0.953	0.000	0.000
Variances						
timedelta	0.000				0.000	0.000
.a3	0.001	0.000	42.076	0.000	0.001	0.000
.a4	0.020	0.000	127.629	0.000	0.020	0.001
.a5	0.002	0.000	66.854	0.000	0.002	0.000
.b1	93.207	52.376	1.780	0.075	93.207	0.726
.b2	6.339	12.678	0.500	0.617	6.339	0.427
Publication	45864.962	325.767	140.791	0.000	1.000	1.000
.Words	12.393	0.088	140.771	0.000	1.000	1.000
Links	35.209	52.375	0.672	0.501	1.000	1.000
R-Square						
timedelta	1.000					
a3	1.000					
a4	0.999					
a5	1.000					
b1	0.274					
b2	0.573					
Words	0.000					

The results in Table 9 reveal the regressions, covariance, variances, and r-square of both latent variables and observed variables. The standard deviation as a whole of the regression of words and links had the value of 0.006. While the covariance of publication and links had 0.000. Further, the variances of observed variables time delta, a3(rate of unique words in the content, and a5 (rate of unique non-stop words in the content) pegged at 0.000. Next was a4 (rate of non-stop words in the content) which had 0.001 while b1 (number of links) had 0.726 and b2 (number of links to other articles) got 0.427. Meanwhile, publication, words, and links which were the latent variables acquired 1.000.

4. Structural Equation Model Derived from the Mined Data



Model 1. Confirmatory Factor Analysis



Model 2. Structural Equation Model

In both confirmatory factor analysis and structural equation model, the publication of articles, words and links served as latent variables while time delta, a3 (Rate of unique words in the content), a4 (Rate of non-stop words in the content), a5 (Rate of unique non-stop words in the content), b1 (Number of links), and b2 (Number of links to other articles) served as the observed variables.

Figure 1 and 2 shows the regressions of the latent variables to all observed variables as well as the covariance of latent variables.

5. Relationship between Confirmatory Factor Analysis and the Structural Equation Model

There was a relationship between the confirmatory factor analysis and the structural equation model. Confirmatory Factor Analysis specifies the measurement model before looking at the data or what is called 'the no peeking rule.' It also tells the indicators measure which factors or unrelated to which factors, whether correlated or uncorrelated. Meanwhile, the structural equation integrates several different multivariate techniques into one fitting model. The results of SEM were integrated with the path analysis using lavaan and regression modeling. It focuses on the indirect as well as the direct effects of variables on other variables.

CONCLUSIONS

The analysis of the mined data from the Mashable website gathered in two years consisted of a heterogeneous set of features which is extracted into 61 attributes and from a total of 39,644 respondents. Seven factors are suggested that existed in the Principal Component Analysis and that it is sufficient to test the hypothesis. On the other hand, the path analysis presents a good model based on the results of CFI, TLI, RMSEA, and SRMR. The researcher also used the confirmatory factor analysis using the R studio to verify the scales used to assess the validity of both convergent and discriminatory. With this, Comparative Fit Index (CFI) with the result of 0.989, Tucker-Lewis Index (TLI) which pegged at 0.979, Root Mean Square Error Estimation (RMSEA) acquired 0.148, and Standardized Root Mean Square Residual (SRMR) has 0.014 which tells the goodness of the model. The researcher also created a structural equation model. The results of SEM are integrated with the path analysis using lavaan and regression modeling. It focuses on the indirect as well as the direct effects of variables on other variables.

RECOMMENDATIONS

In future studies, the researcher may compare and analyze the mined data based on the other variables and may use this study as their reference guide. They may also use other data with a smaller number of variables to have good results in terms of the fitness of the model. Research in the same context which utilizes qualitative methodology is also encouraged.

REFERENCES

Choudhary, S., Sandhu, A. S., & Pradhan, T. (2017). Genetic algorithm-based correlation enhanced prediction of online news popularity. In *Computational Intelligence in Data Mining* (pp. 133-144). Springer, Singapore.

Fernandes, K., Vinagre, P., & Cortez, P. (2015, September). A proactive intelligent decision support system for predicting the popularity of online news. In Portuguese Conference on Artificial Intelligence (pp. 535-546). Springer, Cham.

Gandrud, C. (2013). *Reproducible research with R and R studio*. CRC Press.

Garson, G. D. (2013). *Path analysis*. Asheboro, NC: Statistical Associates Publishing.

Guan, X., Peng, Q., Li, Y., & Zhu, Z. (2017, October). Hierarchical neural network for online news popularity prediction. In 2017 Chinese Automation Congress (CAC) (pp. 3005-3009). IEEE.

Kronthaler, F., & Zöllner, S. (2021). Data Analysis with RStudio. Springer: Berlin/Heidelberg, Germany.

Loraine, A. E., Blakley, I. C., Jagadeesan, S., Harper, J., Miller, G., & Firon, N. (2015). Analysis and visualization of RNA-Seq expression data using RStudio, Bioconductor, and Integrated Genome Browser. In Plant Functional Genomics (pp. 481-501). Humana Press, New York, NY.

Rathord, P., Jain, A., & Agrawal, C. (2019). A comprehensive review on online news popularity prediction using machine learning approach. *trees*, *10*(20), 50.

Ren, H., & Yang, Q. (2015). Predicting and evaluating the popularity of online news. *Standford University Machine Learning Report*.

Wicaksono, A. S., & Supianto, A. A. (2018). Hyper parameter optimization using genetic algorithm on machine learning methods for online news popularity prediction. *International Journal of Advanced Computer Science and Applications*, *9*(12), 263-267. https://archive.ics.uci.edu/ml/datasets/online+news+popularity

AUTHOR'S PROFILE

John Michael D. Aquino He graduated from Southern Luzon State University-Main Campus with Bachelor of Secondary Education Major in MAPEH (Batch 2015) and he passed the Licensure Examination for Teachers in the same year. He also graduated with Masters of Education Major in Physical Education degree at Laguna State Polytechnic University Sta. Cruz Campus. He is now taking up Doctor of Philosophy in Educational Leadership and Management at Philippine Normal University-Manila. He is a former DepEd MAPEH teacher at Division of Quezon and currently teaching at Laguna State Polytechnic University Sta. Cruz Campus as a Full-time P.E Faculty. He is currently the Sports, Spiritual, Cultural Affairs, and Health and Wellness Coordinator in the College of Industrial Technology.

COPYRIGHTS