A Project Report on

ANALYSING ONLINE NEWS POPULARITY: UNVEILING DYNAMICS THROUGH REGRESSION MODELS

Submitted to

Maratha Vidya Prasarak Samaj’s

K.R.T Arts, B.H. Commerce and A.M. Science (K.T.H.M.) College, Nashik-422002

Affiliated to Savitribai Phule Pune University, Pune

Submitted by,

Chanchal Laxman Kotkar

Sanvedana Suryakant Patil

Tejal Sanjay Londhe

Under the guidance of

Dr. Nutan V. Khangar,

Assistant Professor,

Department of Statistics,

K.T.H.M. College, Nashik

May 2023

CERTIFICATE

This is to certify that the project entitled **“Analysing Online News Popularity: Unveiling Dynamics Through Regression Models”** is being submitted by Kotkar Chanchal Laxman, Patil Sanvedana Suryakant, Londhe Tejal Sanjay as partial fulfilment for the award of the degree of the Master of Science (Statistics).

This is a record of considerable work carried out by them under my supervision and guidance.

Place: Nashik

Date:

Project Guide	**Head,**	**Examiner**
Dr. N. V. Khangar	Dr. G. S. Phad	

Department of Statistics
K.T.H.M. College,
Nashik

ACKNOWLEDGEMENT

We have satisfaction on completion of our project entitled “Analysing Online News Popularity: Unveiling Dynamics Through Regression Models” at Department of Statistics in M.V.P. Samaj’s K.R.T. Arts, B.H. Commerce, A.M. Science College, Nashik during the academic year of 2022-2023.

We especially thank the university for providing us this golden opportunity to work on this project. Also, we thank Dr. G. S. Phad, Head, Department of Statistics, K.T.H.M. College, Nashik for providing us access to the department in order to complete our project. Without his support we could not carry out the project successfully.

We take this opportunity to express our deepest gratitude to Dr. Nutan Khangar for her valuable guidance and help. She provided in the due course of project completion. Also, we would like to thank her for helping us in each step of our project analyses.

We are thankful to all faculties of the Department of Statistics for their valuable guidance and great co-operation. Also, I thank non-teaching staff of Department of Statistics, K.T.H.M. College, for cooperating and helping us in the due course of our project completion.

We also want to thank our friends and family who supported us throughout and motivated us to fulfil our project.

Chanchal Laxman Kotkar
Sanvedana Suryakant Patil
Tejal Sanjay Londhe

ABSTRACT

In this project, we study online dissemination of news on the Mashable website. Here our main aim is to extract relevant variables that are useful for forecasting the number of news shares. The related data is obtained through the news articles which is on UCI Machine Learning Repository. The link of the dataset is https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity .

In this, we have performed Poisson regression for under-dispersion variables, negative binomial for over-dispersion variables and geometric regression was performed to resolve the both issues of under dispersion and over dispersion. After fitting all the regression models, it can be concluded that limitations of Poisson and Negative Binomial regression can be solved by Geometric regression. Using Poisson regression, only under dispersion problem can be solved. Poisson model appeared to outperform than other two models. Negative Binomial regression can only solve the problem of overdispersion. So, both Poisson regression and Negative Binomial regression cannot solve the under dispersion and over dispersion problem simultaneously.

Limitations of Poisson and Negative Binomial regression can be solved by Geometric regression. Geometric model can be used as an alternative modelling approach for both under-dispersed and over-dispersed count data. So, if the data is count then one can use the Geometric regression model to solve under-dispersion and over-dispersion problem simultaneously and using regression line equation, one can predict the count of shares of news. So, Geometric model may serve as an alternative model to Poisson and NB models for modelling count data.

INDEX

1. INTRODUCTION

News is the most important part of our routine life. Nowadays everyone prefers reading news on online platforms. There are lot of platforms such as Google News, Daily Hunt, TOI, Inshorts and one of them is Mashable.

Mashable is a news website, digital media platform and entertainment company founded by Pete Cashmore while living in Aberdeen, Scotland, in July 2005. It is a global, multi-platform media and entertainment company, powered by its own proprietary technology. Mashable is the go-to source for tech, digital culture, and entertainment content for its dedicated and influential audience around the globe. Early iterations of the site were a simple WordPress blog, with Cashmore as sole author. Fame came relatively quickly, with Time magazine noting Mashable as one of the 25 best blogs of 2009. As of November 2015, it had over 6,000,000 Twitter followers and over 3,200,000 fans on Facebook. In June 2016, it acquired YouTube channel CineFix from Whalerock Industries.

News share data is the count data. So, to analyze this type of data we can use various regression models such as Poisson regression, negative binomial regression, geometric regression. Also, we can use nonlinear factor analysis as a statistical method to analyze count data. We can also use different Machine Learning models of classification and regression to analyze count data.

Yalcin and Amemiya (2001) use Nonlinear Factor Analysis as a Statistical method to analyze count data. Roja Bandari et al. (2010-11) evaluated multi-dimensional feature space derived from properties of an article is constructed and the efficacy of these features to serve as predictors of online popularity. Both regression and classification algorithms are examined and demonstration that despite randomness in human behavior. Alexandru Tatar et al. (2012) addressed the problem of predicting the popularity of news articles based on user comments. Gabor Szabo and Bernardo A. Huberman (2008), using two content sharing portals, YouTube and Digg, showed that by modelling the accrual of views and votes on content offered by these services we can predict the long-term dynamics of individual submissions from initial data. Sasa Petrovic et al. (2011) predicting if a tweet will be retweeted, and solving this problem furthers our understanding of message propagation within large user communities.

Data for this project contained some under dispersed variables and over dispersed variables. Poisson regression is used for under-dispersion variables, negative binomial for over-dispersion variables and geometric regression was performed to resolve the both issues of under dispersion and over dispersion. Harris et al. (2012) introduced a supporting Stata program and illustrated the effectiveness of three Poisson regression models (Poisson, GP, and QP) when dealing with under-dispersed count data and compare it using simulation study. To solve the over dispersion problem one can, use Negative Binomial regression. Linde and Mantyniemi (2011) proposed parameterization of negative binomial dispersion and illustrated it by applying the model to empirical migration data with a high level of dispersion. In some cases, the variables are both under-dispersed and over- dispersed. So, to solve both the problems simultaneously one can use Geometric Regression model. Al-balushi and Islam (2020) studied Geometric regression for modelling count data and compared Geometric, Poisson and Negative Binomial regression model to check goodness of fit.

The data was obtained from UCI Machine Learning Repository. This dataset summarizes a heterogenous set of features about articles published by Mashable in a period of two years. There are various variables like number of words in the title as well as description, best or worst keyword, number of shares of those keywords, shares of referenced articles, type of news, day of publishing, global subjectivity, global sentiment polarity, number of shares of the news, etc. The of this dataset is *(https://archive.ics.uci.edu/dataset/332/online+news+popularity)*

Here, our aim is to predict the popularity of a news article by taking into consideration all the concerned variables. The goal of our project is to determine the impact of various variables on the revenue of the website.

1.1 Literature Review

In order to have clarity in the analysis we went through a lot of research papers to serve the purpose. The abstract description of them is given below:

Yalcin and Amemiya (2001) reviewed the statistical contributions to the issues like identification ambiguity and heavy reliance. They have studied limitation to linearity in detail. Gabor Szabo and Bernardo A. Huberman (2008), using two content sharing portals, YouTube and Digg, showed that by modelling the accrual of views and votes on content offered by these services we can predict the long-term dynamics of individual submissions from initial data. The differing time scales of the predictions are shown to be due to differences in how content is consumed on the two portals: Digg stories quickly become outdated, while YouTube videos are still found long after they are initially submitted to the portal. Roja Bandari et al. (2010-11) evaluated multi-dimensional feature space derived from properties of an article is constructed and the efficacy of these features to serve as predictors of online popularity. Both regression and classification algorithms are examined and demonstration that despite randomness in human behaviour, it is possible to predict ranges of popularity on twitter with an overall 84% accuracy is done. Sasa Petrovic et al. (2011) predicting if a tweet will be retweeted, and solving this problem furthers our understanding of message propagation within large user communities. A human experiment on the task of deciding whether a tweet will be retweeted which shows that the task is possible is carried out, as human performance levels are much above chance. Alexandru Tatar et al. (2012) addressed the problem of predicting the popularity of news articles based on user comments. Prediction task as a ranking problem is formulated, where the goal is not to infer the precise attention that a content will receive but to accurately rank articles based on their predicted popularity. Using data obtained from two important news sites in France and Netherlands, the ranking effectiveness of two prediction models is analysed.

Linde and Mantyniemi (2011) proposed parameterization of negative binomial dispersion and illustrated it by applying the model to empirical migration data with a high level of dispersion. Modelling under dispersed count data with generalized Poisson regression introduced by Harris et al. (2012). They introduced a supporting Stata program and illustrated the effectiveness of three Poisson regression models (Poisson, GP, and QP) when dealing with under-dispersed count data and compare it using simulation study. Al-balushi and Islam (2020) studied Geometric regression for modelling count data and compared Geometric, Poisson and Negative Binomial regression model to check goodness of fit.

2. MOTIVATION

Being a M.Sc. II Statistics, we were always eager to know how the big data is handled. Today we are living in the e-world where every individual is connected to each other by the means of internet. Also, we understand news is an inseparable part of our life. Nowadays everyone prefers to scroll news on online news platforms rather than traditional newspapers. Considering the above scenario, we focused on taking the data about all news articles published on an online news platform "Mashable" in the period of two years.

This was the greatest motivation for us to work on the above data, in order to predict the popularity of a particular news. This will ultimately help us in optimization of news shares over the platform if we already have an idea about the significant factors influencing the popularity of the news.

With this approach we tend to work on the task that would help the platform to convey complete and authentic information to its end users. This is also one of the major reasons that encouraged us to choose this data.

3. OBJECTIVES

There are various objectives for this project such as to find the significant factors affecting the popularity of the news which in turn can be used for the optimization of revenue by the company. Also, to visualize about the different aspects of the news popularity. In this project, our aim is to study the significance of the factors responsible for the popularity of the news. Also, to predict the number of shares of news based on concerned variables under study. For this analysis, we develop the code in R/ Python whenever needed.

Conversion of Technical Problem to statistical problem following statistical methodologies:

In order to fulfil the objectives mentioned above we apply several statistical techniques to dataset. This helps us reach our goal of obtaining results on the new popularity. These techniques enlisted here. To study the cases of equi-dispersion, over-dispersion, and under-dispersion, we overviewed the different regression models for the count dataset such as Poisson regression, Negative Binomial regression, Geometric regression. Further, we compare goodness of fit of Geometric, Poisson and Negative Binomial regression model numerically. Also, we find out the significant variables which are significantly affects the popularity of news.

4. METHODOLOGY

A sort of regression analysis in which data is fitted to a model and then displayed numerically is known as nonlinear regression. Simple linear regression connects two variables (X and Y) in a straight line (y = mx + b), whereas nonlinear regression connects two variables (X and Y) in a nonlinear (curved) relationship. The goal of the model is to minimise the sum of squares as much as possible. The sum of squares is a statistic that tracks how much Y observations differ from the nonlinear (curved) function that was used to anticipate Y.

In the same way that linear regression modelling aims to graphically trace a specific response from a set of factors, nonlinear regression modelling aims to do the same. Because the function is generated by a series of approximations (iterations) that may be dependent on trial-and-error, nonlinear models are more complex to develop than linear models.

For the count data set the following regression models are used widely. The brief discussion is given in the following subsections.

4.1 Poisson Regression

The Poisson distribution models the probability of y events (i.e., failure, death, or existence) with the formula,

$$P(Y = y|\mu) = \frac{e^{-\mu}\mu^{y}}{y!}, y = 0,1,2,\ldots$$

The Poisson distribution is specified with a single parameter μ. This is the mean incidence rate of a rare event per unit of exposure. Exposure may be time, space, distance, area, volume, or population size. Because exposure is often a period of time, we use the symbol t to represent the exposure. When no exposure value is given, it is assumed to be one. The parameter μ may be interpreted as the risk of a new occurrence of the event during a specified exposure period, t. The probability of y events is then given by

$$P(Y = y|\mu, \mathrm{t}) = \frac{e^{-\mu t}(\mu t)^{y}}{y!}, \quad y = 0, 1, 2, \ldots$$

where, $E(Y) = Var(Y)$

- **The Poisson Regression Model**

In Poisson regression, the Poisson incidence rate μ is determined by a set of k regressor variables. The expression relating these quantities is

$$\mu \;=\; te^{(\beta_1 X_1 \,+\, \beta_2 X_2 \,+\, \cdots \,+\, \beta_k X_k)}$$

where, X_1 : 1

β_1 : intercept

β_j : unknown parameters ; $j = 2, 3, \dots, k$

The fundamental Poisson regression model for an observation i is written as

$$P(Y = y_i | \mu_i, t_i) = \frac{e^{-\mu_i, t_i}(\mu_i, t_i)^{y_i}}{y_i!}$$

where, $\mu_i = t_i \mu(X_i'\beta) = t_i e^{(\beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})}$

- **Parameter Estimation using Maximum Likelihood Estimation for Poisson Regression Model**

The regression coefficients are estimated using the method of maximum likelihood. The logarithm of the likelihood function,

$$ln[L(y,\beta)] = \sum_{i=1}^{n} y_i \, ln[t_i \mu(X_i'\beta)] \; - \sum_{i=1}^{n} t_i \mu(X_i'\beta) - \sum_{i=1}^{n} \ln{(y_i!)}$$

This will make their calculated log-likelihoods different from ours. The likelihood equations may be formed by taking the derivatives with respect to each regression coefficient and setting the result equal to zero. Doing this leads to a set of nonlinear equations that admits no closed form solution. Thus, an iterative algorithm must be used to find the set of regression coefficients that maximum the log-likelihood. Using the method of iteratively reweighted least squares, a solution may be found in five or six iterations. However, the algorithm requires a complete pass through the data at each iteration, so it is relatively slow for problems with a large number of rows. With today's computers, this is becoming less and less of an issue.

For various reasons the amount of variation for each sampling unit is typically higher than expected by a pure Poisson process. We can't solve the Overdispersion problem by using simple poisson regression. The above information is collected from NCSS report of Poisson regression (*https://rb.gy/gsmle*). To solve the problem of Overdispersion we use negative binomial approach. The detailing of the negative binomial regression is given in the next subsection.

4.2 Negative Binomial Regression

The Poisson distribution may be generalized by including a gamma noise variable which has a mean of 1 and a scale parameter of v. The Poisson-gamma mixture (negative binomial) distribution that results is

$$P(Y = y_i|\mu_{i,} \alpha) = \frac{|\overline{(y_i + \alpha^{-1})}}{|\overline{(y_i + 1)}|\overline{\alpha^{-1}}}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}} \left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{y_i}$$

where, $\mu_i = t_i\mu$ and $\alpha = 1/v$

The parameter μ is the mean incidence rate of y per unit of exposure. Exposure may be time, space, distance, area, volume, or population size. Because exposure is often a period of time, we use the symbol t_i to represent the exposure for a particular observation. When no exposure given, it is assumed to be one. The parameter μ may be interpreted as the risk of a new occurrence of the event during a specified exposure period, t.

The results below make use of the following relationship derived from the definition of the gamma function

$$ln\left(\frac{|\overline{(y_i + \alpha^{-1})}}{|\overline{\alpha^{-1}}}\right) = \sum_{j=0}^{y_i-1} \ln{(j + \alpha^{-1})}$$

- **The negative binomial regression model:**

In negative binomial regression, the mean of y is determined by the exposure time t and a set of k regressor variables (the $X's$). The expression relating these quantities is

$$\mu_i = \mathrm{e}^{(\ln(t_i)+\beta_1 X_{1i}+\beta_2 X_{2i}+\cdots+\beta_k X_{ki})}$$

where, $X_1 : 1$

β_1 : intercept

β_j : unknown parameters ; $j = 2, 3, \dots, k$

The fundamental Poisson regression model for an observation i is written as

$$P(Y = y_i|\mu_{i,}\alpha) = \frac{\overline{|(y_i + \alpha^{-1})}}{\overline{|(y_i + 1)}\overline{|\alpha^{-1}}}\left(\frac{1}{1 + \alpha\mu_i}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^{y_i}$$

- **Parameter Estimation using Maximum Likelihood Estimation for Negative Binomial Regression Model**

The regression coefficients are estimated using the method of maximum likelihood. Cameron gives the logarithm of the likelihood function as

$$L = \sum\nolimits_{i=1}^{n}\Big\{\ln\Big[\overline{|(y_i + \alpha^{-1})}\Big] - ln\Big[\overline{|\alpha^{-1}}\Big] - ln\Big[\overline{|(y_i + 1)}\Big] - \alpha^{-1}\ln(1 + \alpha\mu_i) \dots$$

$$- y_i ln(1 + \alpha\mu_i) + y_i \ln(\alpha) + y_i \ln(\mu_i)\Big\}$$

Rearranging gives,

$$L = \sum\nolimits_{i=1}^{n}\left\{\left(\sum\nolimits_{j=0}^{y_i-1}\ln\,(j + \alpha^{-1})\right) - ln\Big[\overline{|(y_i + 1)}\Big] - (y_i + \alpha^{-1})\ln(1 + \alpha\mu_i) \dots\right.$$

$$\left. + y_i \ln(\alpha) + y_i \ln(\mu_i)\right\}$$

The first derivatives of $\mathscr{L}$ were given by Cameron (2013) and Lawless (1987) as

$$\frac{\partial L}{\partial \beta_j} = \sum\nolimits_{i=1}^{n} \frac{x_{ij}(y_i - \mu_i)}{1 + \alpha\mu_i}, \quad j = 1, 2, \dots, k$$

$$\frac{\partial L}{\partial \alpha} = \sum\nolimits_{i=1}^{n} \left\{ \alpha^{-2} \left(\ln(1 + \alpha\mu_i) - \sum\nolimits_{j=0}^{y_i - 1} \frac{1}{j + \alpha^{-1}} \right) + \frac{y_i - \mu_i}{\alpha(1 + \alpha\mu_i)} \right\}$$

Equating the gradients to zero gives the following set of likelihood equations

$$\sum\nolimits_{i=1}^{n} \frac{x_{ij}(y_i - \mu_i)}{1 + \alpha\mu_i} = 0, \quad j = 1, 2, \dots, k$$

$$\sum\nolimits_{i=1}^{n} \left\{ \alpha^{-2} \left(\ln(1 + \alpha\mu_i) - \sum\nolimits_{j=0}^{y_i - 1} \frac{1}{j + \alpha^{-1}} \right) + \frac{y_i - \mu_i}{\alpha(1 + \alpha\mu_i)} \right\} = 0$$

The above information is collected from NCSS report of negative binomial regression (*https://rb.gy/8lrb8*). We can't solve both issues of under dispersion and over dispersion by Poisson regression and negative binomial regression. Geometric regression was performed to resolve the both issues of under dispersion and over dispersion. The detailing of the geometric regression is given in the next subsection.

4.3 Geometric regression

The Poisson distribution may be generalized by including a gamma noise variable which has a mean of 1 and a scale parameter of v. The Poisson-gamma mixture (negative binomial) distribution that results is

$$P(Y = y_i | \mu_{i,} \alpha) = \frac{\overline{|(y_i + \alpha^{-1})}}{\overline{|(y_i + 1)}\overline{|\alpha^{-1}}} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i} \right)^{\alpha^{-1}} \left(\frac{\mu_i}{\alpha^{-1} + \mu_i} \right)^{y_i}$$

Where, $\mu_i = t_i\mu$ and $\alpha = 1/v$

The parameter μ is the mean incidence rate of y per unit of exposure. Exposure may be time, space, distance, area, volume, or population size. Because exposure is often a period of time, we use the symbol t_i to represent the exposure for a particular observation. When no exposure given, it is assumed to be one. The parameter μ may be interpreted as the risk of a new occurrence of the event during a specified exposure period, t. When the dispersion parameter α is set to one, the result is called the geometric distribution.

- ## The Geometric Regression Model

In geometric regression, the mean of y is determined by the exposure time t and a set of *k* regressor variables (the *x's*). The expression relating these quantities is

$$\mu_i = \mathrm{e}^{(\ln(t_i)+\beta_1 X_{1i}+\beta_2 X_{2i}+\cdots+\beta_k X_{ki})}$$

where, $X_1 : 1$

β_1 : intercept

β_j : unknown parameters ; $j = 2, 3, \ldots, k$

The fundamental geometric regression model for an observation i is written as

$$P(Y = y_i|\mu_i) = \frac{\overline{|(y_i + 1)}}{\overline{|(y_i + 1)}}\left(\frac{1}{1+\mu_i}\right)^1 \left(\frac{\mu_i}{1+\mu_i}\right)^{y_i}$$

- ## Parameter Estimation using Maximum Likelihood Estimation for Geometric Regression Model

The regression coefficients are estimated using the method of maximum likelihood. Cameron gives the logarithm of the likelihood function as

$$L = \sum\nolimits_{i=1}^{n} \left\{\ln\left[\overline{|(y_i + 1)}\right] - ln\left[\overline{|(y_i + 1)}\right] - \ln(1+\mu_i) - y_i ln(1+\mu_i) + \mathrm{y}_i \ln\,(\mu_i)\right\}$$

Rearrangement of terms gives

$$L = \sum\nolimits_{i=1}^{n} \left\{ \sum\nolimits_{j=0}^{y_i - 1} ln(j+1) - \ln\ |\overline{(y_i + 1)} - (y_i + 1) ln(1 + \mu_i) + y_i ln(\mu_i) \right\}$$

The first derivatives of $\mathscr{L}$ were given by Cameron (2013) and Lawless (1987) as

$$\frac{\partial L}{\partial \beta_j} = \sum\nolimits_{i=1}^{n} \frac{x_{ij}(y_i - \mu_i)}{1 + \mu_i}, \qquad j = 1, 2, \ldots, k$$

$$\frac{-\partial^2 L}{\partial \beta_r \partial \beta_s} = \sum\nolimits_{i=1}^{n} \frac{\mu_i (1 + y_i) x_{ir} x_{is}}{(1 + \mu_i)^2}, \qquad r, s = 1, 2, \ldots, k$$

Equating the gradients to zero gives the following set of likelihood equations

$$\sum\nolimits_{i=1}^{n} \frac{x_{ij}(y_i - \mu_i)}{1 + \mu_i} = 0, \qquad j = 1, 2, \ldots, k$$

$$\sum\nolimits_{i=1}^{n} \left\{ \left(\ln(1 + \mu_i) - \sum\nolimits_{j=0}^{y_i - 1} \frac{1}{j+1} \right) + \frac{y_i - \mu_i}{(1 + \mu_i)} \right\} = 0$$

The above information is collected from NCSS report of geometric regression (*https://rb.gy/uno0n*).

When we study above three regression models comparatively, we observed that we can solve the different dispersion problems viz. under-dispersion and over-dispersion problems using above regression models like we can solve the under-dispersion problem using poisson regression whereas by using negative binomial regression we can solve the problem of over-dispersion and both the problems can solve by using geometric regression at a time.

5. CASE STUDY: ONLINE NEWS POPULARITY

We obtained this dataset from the UCI Machine Learning Repository. This dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of two years. It consists of 61 attributes which describe the popularity of the article. It includes various factors like number of words in the title as well as the description, best or worst keyword, number of shares of those keywords, shares of referenced articles, day of publishing, global subjectivity, global sentiment polarity etc. This is the data we finalized for our analysis considering the abundance of attributes and the insights they display.

Data Description

Link: *https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity*

Dataset Name: Online News Popularity Data Set

Description of Variables:

Number of Variables: 42 (41 independent variables, 1 dependent variable)

Attribute Information:

X_1: n_tokens_title: Number of words in the title

X_2: num_keywords: Number of keywords in the metadata

X_3: n_tokens_content: Number of words in the content

X_4: n_unique_tokens: Rate of unique words in the content

X_5: n_non_stop_words: Rate of non-stop words in the content

X_6: n_non_stop_unique_tokens: Rate of unique non-stop words in the content

X_7: num_hrefs: Number of links

X_8: num_self_hrefs: Number of links to other articles published by Mashable

X_9: num_imgs: Number of images

X_{10}: num_videos: Number of videos

X_{11}: average_token_length: Average length of the words in the content

X_{12}: kw_min_min: Worst keyword (min. shares)

X_{13}: kw_max_min: Worst keyword (max. shares)

X_{14}: kw_avg_min: Worst keyword (avg. shares)

X_{15}: kw_min_max: Best keyword (min. shares)

X_{16}: kw_avg_max: Best keyword (avg. shares)

X_{17}: kw_min_avg: Avg. keyword (min. shares)

X_{18}: kw_max_avg: Avg. keyword (max. shares)

X_{19}: kw_avg_avg: Avg. keyword (avg. shares)

X_{20}: self_reference_min_shares: Min. shares of referenced articles in Mashable

X_{21}: self_reference_max_shares: Max. shares of referenced articles in Mashable

X_{22}: self_reference_avg_sharess: Avg. shares of referenced articles in Mashable

X_{23}: is_weekend : Whether the news is published on a weekend

X_{24}: weekdays: Whether the news is published on a weekdays

X_{25}: News_types: The type of the news being published

X_{26}: global_subjectivity: Text subjectivity

X_{27}: global_sentiment_polarity: Text sentiment polarity

X_{28}: global_rate_positive_words: Rate of positive words in the content

X_{29}: global_rate_negative_words: Rate of negative words in the content

X_{30}: rate_positive_words: Rate of positive words among non-neutral tokens

X_{31}: rate_negative_words: Rate of negative words among non-neutral tokens

X_{32}: avg_positive_polarity: Avg. polarity of positive words

X_{33}: min_positive_polarity: Min. polarity of positive words

X_{34}: max_positive_polarity: Max. polarity of positive words

X_{35}: avg_negative_polarity: Avg. polarity of negative words

X_{36}: min_negative_polarity: Min. polarity of negative words

X_{37}: max_negative_polarity: Max. polarity of negative words

X_{38}: title_subjectivity: Title subjectivity

X_{39}: title_sentiment_polarity: Title polarity

X_{40}: abs_title_subjectivity: Absolute subjectivity level

X_{41}: abs_title_sentiment_polarity: Absolute polarity level

X_{42}: shares: Number of shares (target)

5. GRAPHICAL PRESENTATION

Fig. 1 Bar graph of percentage sharing of news by news_type

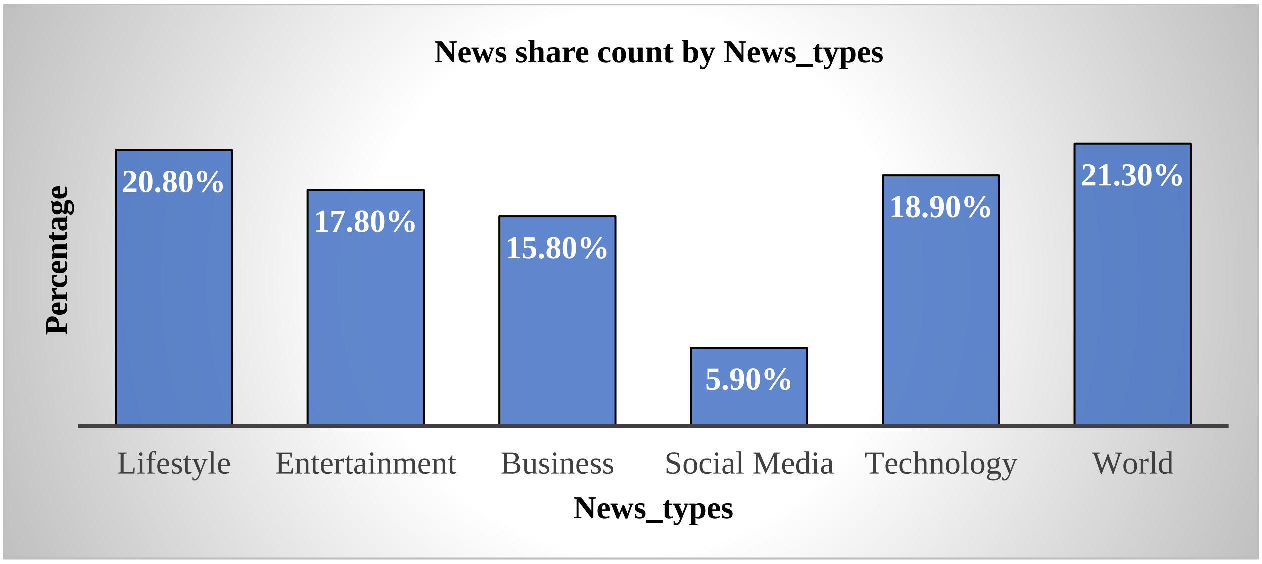
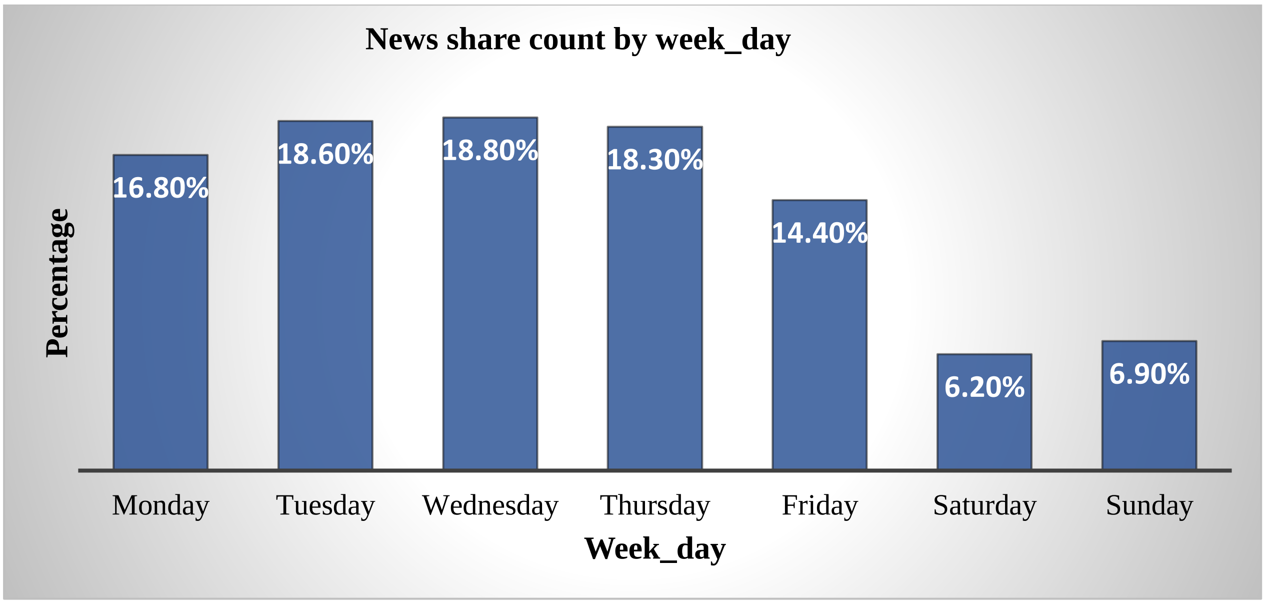
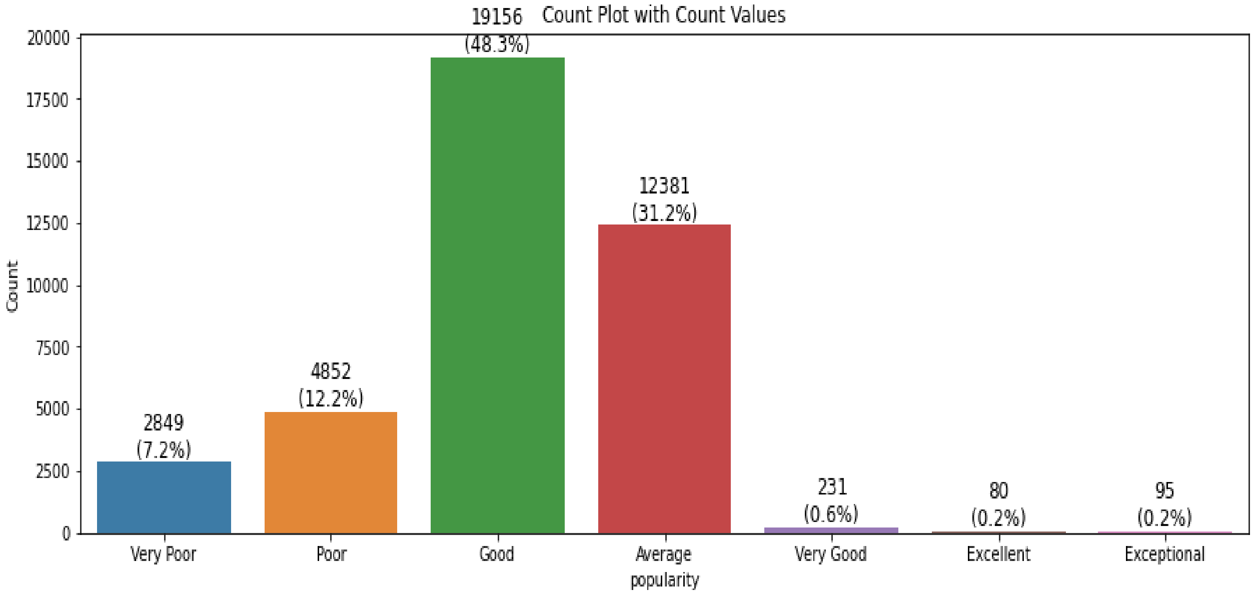


Fig. 2 Bar graph of percentage sharing of news share count by week_day



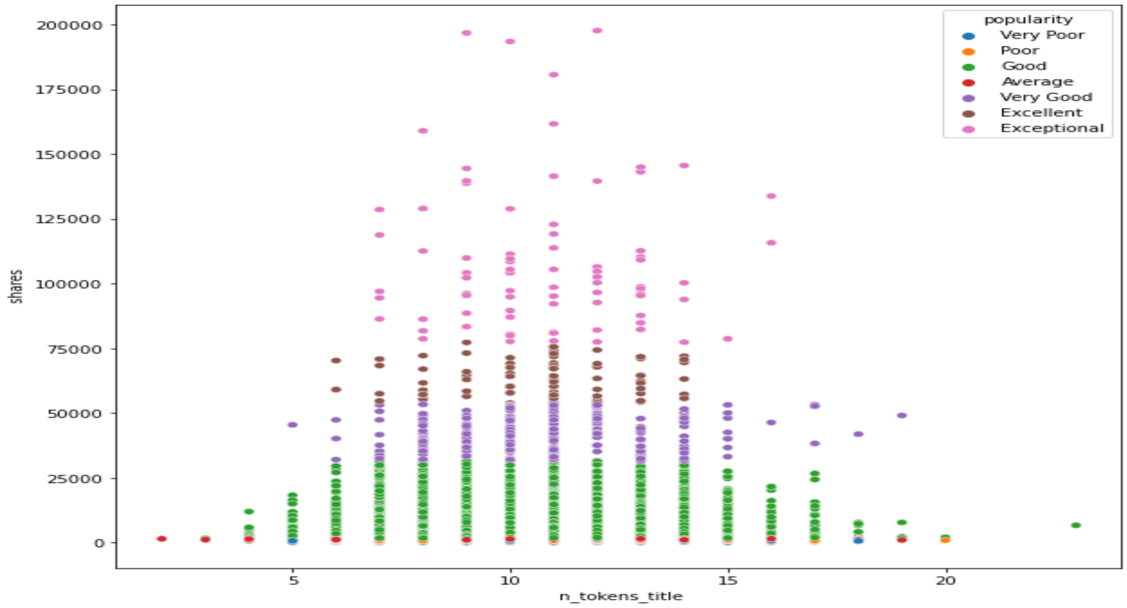
From Fig. 1, it is observed that News articles which are related to world and lifestyle are mostly shared. While news articles related to social media are shred too less i.e., only around 6%. From Fig. 2, it is observed that most of the news are shared on Tuesday, Wednesday, and Thursday. Very few news is shared on Saturday and Sunday. i.e., Mostly news is shared on weekdays not on weekend. Fig.1 and Fig.2 are taken from MS Excel.

Fig. 3 Count plot of popularity of news shared



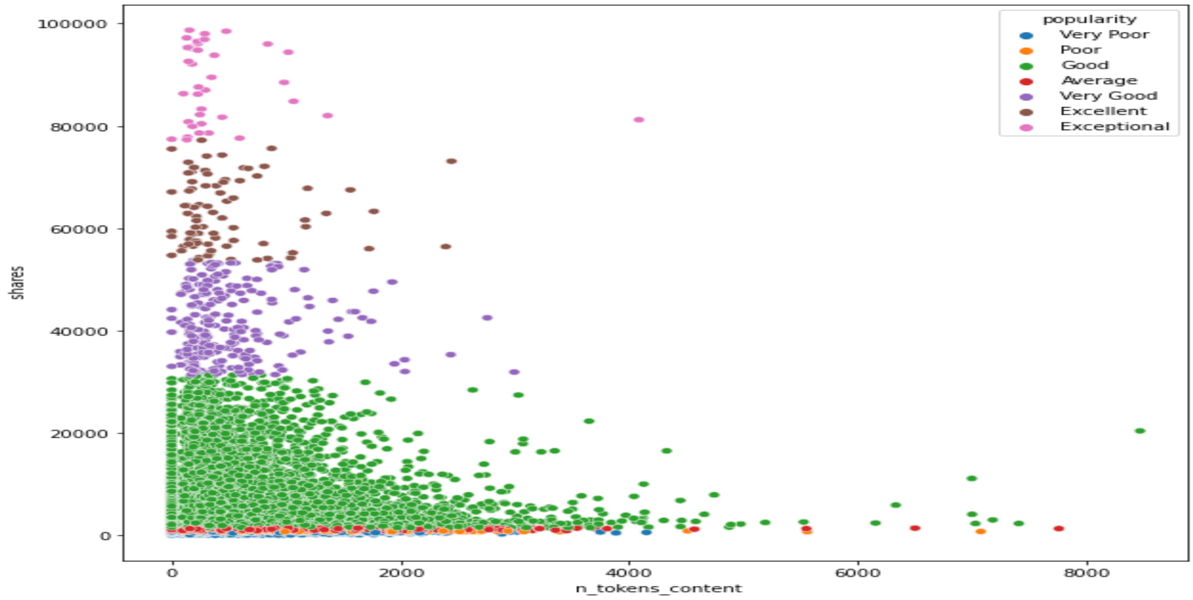
If news share is less than 645 then we say it Very Poor. If news share is between 645 to 861 then we say it Poor. If news share is greater than 861 and less than 1400 (i.e., Top 50%) then news share is Average. If news share is greater than 1400 and less than 31300 then news share is Good. If news share is greater than 31300 and less than 53700 then news share is Very Good. If news share is greater than 53700 and less than 77200 then news share is Excellent. Other than this are Exceptional. From Fig.3 it is observed that 7.2% news are very poorly shared. Very few news is highly shared. (i.e., popular). Nearly 48.3% news are shared between 1400-31300.

Fig. 4 Scatter plot of n_token_title (Number of words in the title) Vs news shares



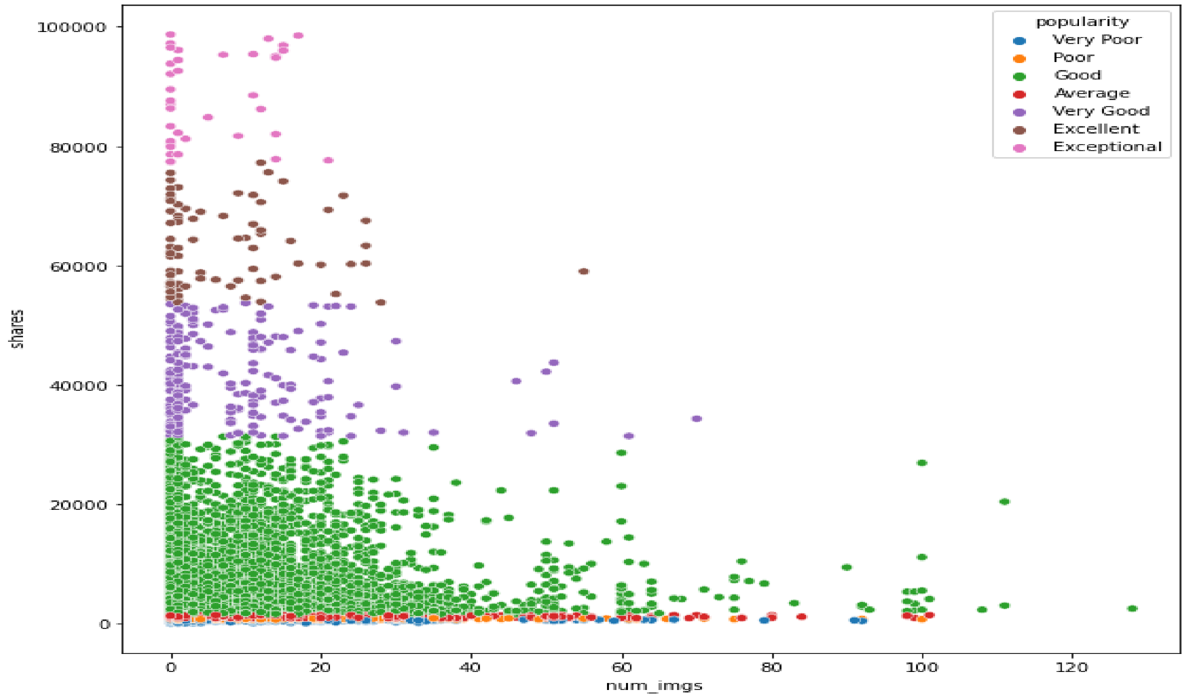
From Fig.4, it is observed that article title shouldn't be too long or too short. Most of the points in the scattered diagram are between 6 - 17. So, 6 - 17 words are the ideal number of words to have for titles.

Fig. 5 Scatter plot of n_token_contents (Number of words in the content) Vs news shares



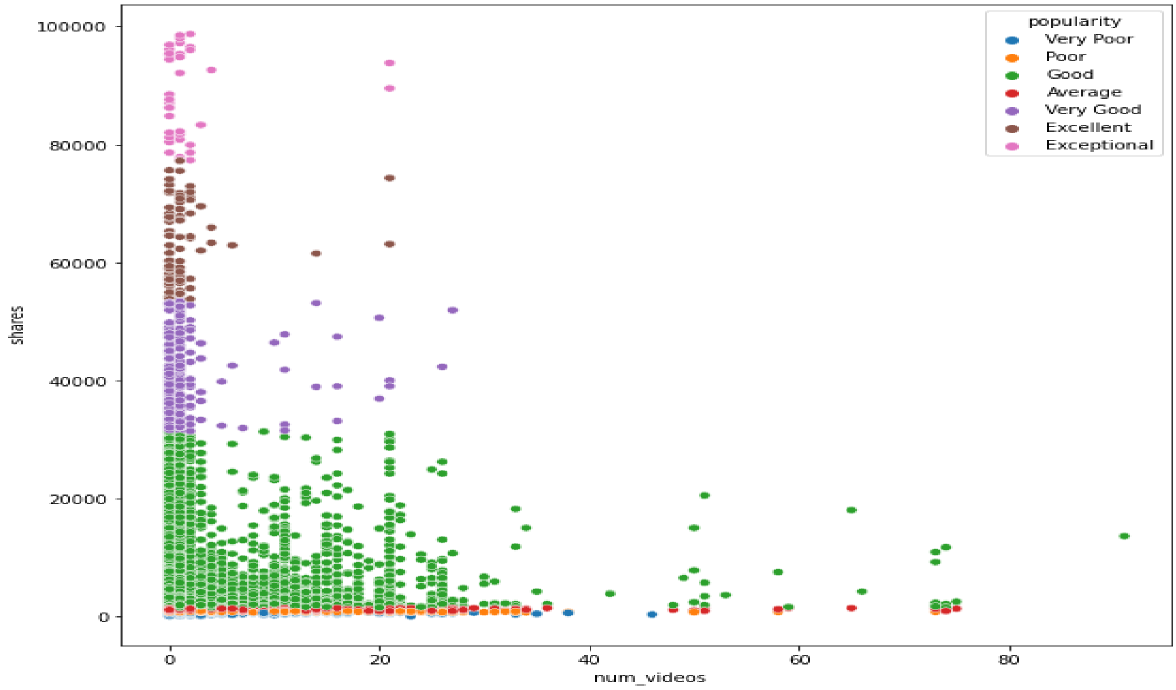
From Fig. 5, it can be observed that most of the points in the scattered diagram are between 0 – 1600. So, the number of words in the article should be less than 1600 words. The lesser the better.

Fig. 6 Scatter plot of num_imgs (Number of images in the content) Vs news shares



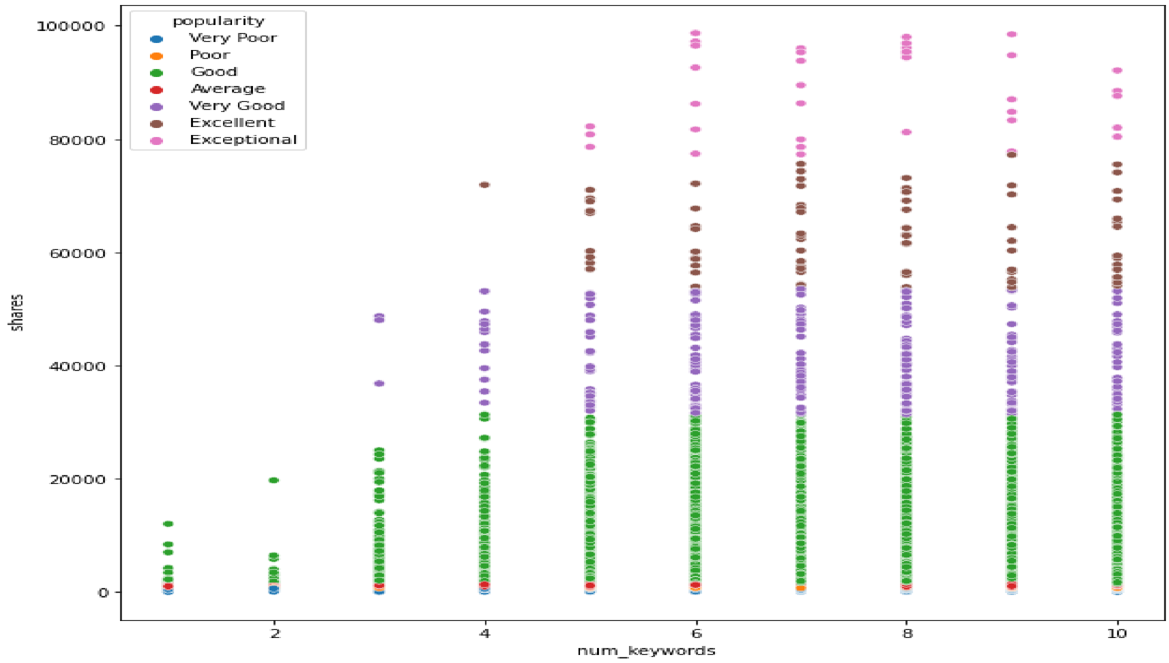
From Fig.6, most of the scatter points are between 1-40. So, articles should have good number of images. Between 1 – 40 images are great.

Fig. 7 Scatter plot of num_videos (Number of videos in the content) Vs news shares



From Fig.7, it is observed that number of videos should be between 0 - 25 videos.

Fig. 8 Scatter plot of num_keywords (Number of images in the content) Vs news shares



From Fig.8, number of keywords in the metadata really influences the shares to a margin. The higher the value the better the shares chances. A value upward of 5 is recommend.

From Fig.3 to Fig.8 all scatter plots are taken from python.

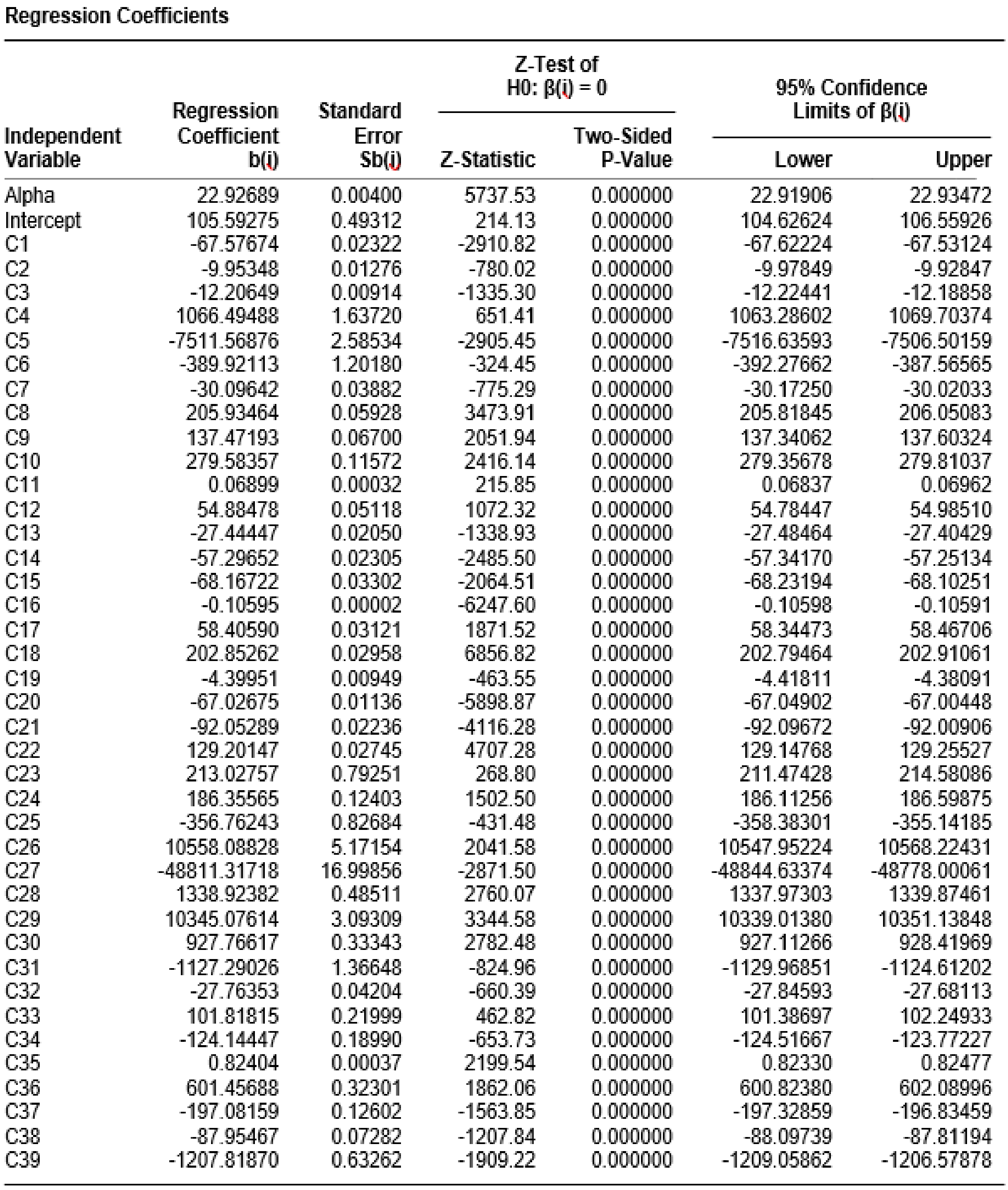
6. NUMERICAL RESULTS

Table 1: Numerical Results Obtain by Performing Poisson Regression Model

Coefficients	Estimate	Std. Error	z value	P(>|z|)
(Intercept)	5.700e+00	1.651e-03	3452.88	<2e-16
n_tokens_title	4.108e-02	8.389e-05	489.62	<2e-16
num_keywords	5.872e-03	4.585e-05	128.08	<2e-16
n_unique_tokens	-2.993 e-01	5.729e-03	-52.25	<2e-16
n_non_stop_words	1.069 e+00	6.385e-03	167.45	<2e-16
n_non_stop_unique_tokens	-1.977 e-01	4.173e-03	-47.39	<2e-16
num_hrefs	9.874 e-02	1.358e-04	726.89	<2e-16
num_self_hrefs	-2.703 e-01	2.192e-04	-1233.44	<2e-16
num_imgs	2.276 e-02	2.188e-04	103.99	<2e-16
num_videos	4.946 e-01	4.051e-04	1221.03	<2e-16
kw_min_min	1.657 e-01	1.779e-04	931.66	<2e-16
kw_max_avg	-5.735 e-02	9.888e-05	-579.95	<2e-16
kw_avg_avg	6.161 e-02	3.341e-05	1844.02	<2e-16
is_weekend	1.262 e+00	2.682e-03	470.51	<2e-16
global_subjectivity	2.512 e-01	4.237e-04	592.91	<2e-16
global_sentiment_polarity	1.714 e-01	2.766 e-03	61.96	<2e-16
global_rate_negative_words	9.644 e+00	6.325e-02	152.46	<2e-16
rate_positive_words	-2.022 e-01	1.152e-03	-175.54	<2e-16
rate_negative_words	-1.923 e+00	7.616e-03	-252.45	<2e-16
avg_positive_polarity	-1.938 e-01	1.174e-03	-165.15	<2e-16
min_positive_polarity	-2.502 e+00	4.872e-03	-513.45	<2e-16
max_positive_polarity	1.072 e-02	1.513e-04	70.84	<2e-16
avg_negative_polarity	-5.154e-02	7.717e-04	-66.79	<2e-16
min_negative_polarity	-4.982e-02	6.768e-04	-73.61	<2e-16
title_subjectivity	-1.193e-01	1.129 e-03	-105.63	<2e-16
title_sentiment_polarity	6.817e-02	4.177e-04	163.23	<2e-16
abs_title_subjectivity	6.663e-02	2.455e-04	271.43	<2e-16
abs_title_sentiment_polarity	6.027e-01	2.240e-03	269.01	<2e-16
global_rate_positive_words	-5.463e+00	1.904e-02	-286.83	<2e-16
n_tokens_content	-3.876e-03	3.201e-05	-121.07	<2e-16
average_token_length	-2.906e-04	1.145e-06	-253.82	<2e-16
kw_max_min	-5.929e-03	7.261e-05	-81.65	<2e-16
kw_avg_min	9.663e-03	8.300e-05	116.42	<2e-16
kw_min_max	-5.527e-02	1.154e-04	-478.80	<2e-16
kw_avg_max	-9.014e-06	5.973e-08	-150.93	<2e-16
kw_min_avg	3.010e-02	1.093e-04	275.50	<2e-16
self_reference_min_shares	-4.498e-04	3.384e-05	-13.29	<2e-16
self_reference_max_shares	8.113e-03	7.453e-05	108.86	<2e-16
self_reference_avg_sharess	1.767e-02	8.709e-05	202.84	<2e-16
max_negative_polarity	-2.833e-04	1.327e-06	-213.53	<2e-16

The Poisson regression is fitted in R software. From table 1, we get the estimates, standard errors, z value and p-value. From the table all the p-values are less than 0.05 so we can say that all variables are significant.

Table 2: Numerical Results Obtain by Performing Negative Binomial Regression Model

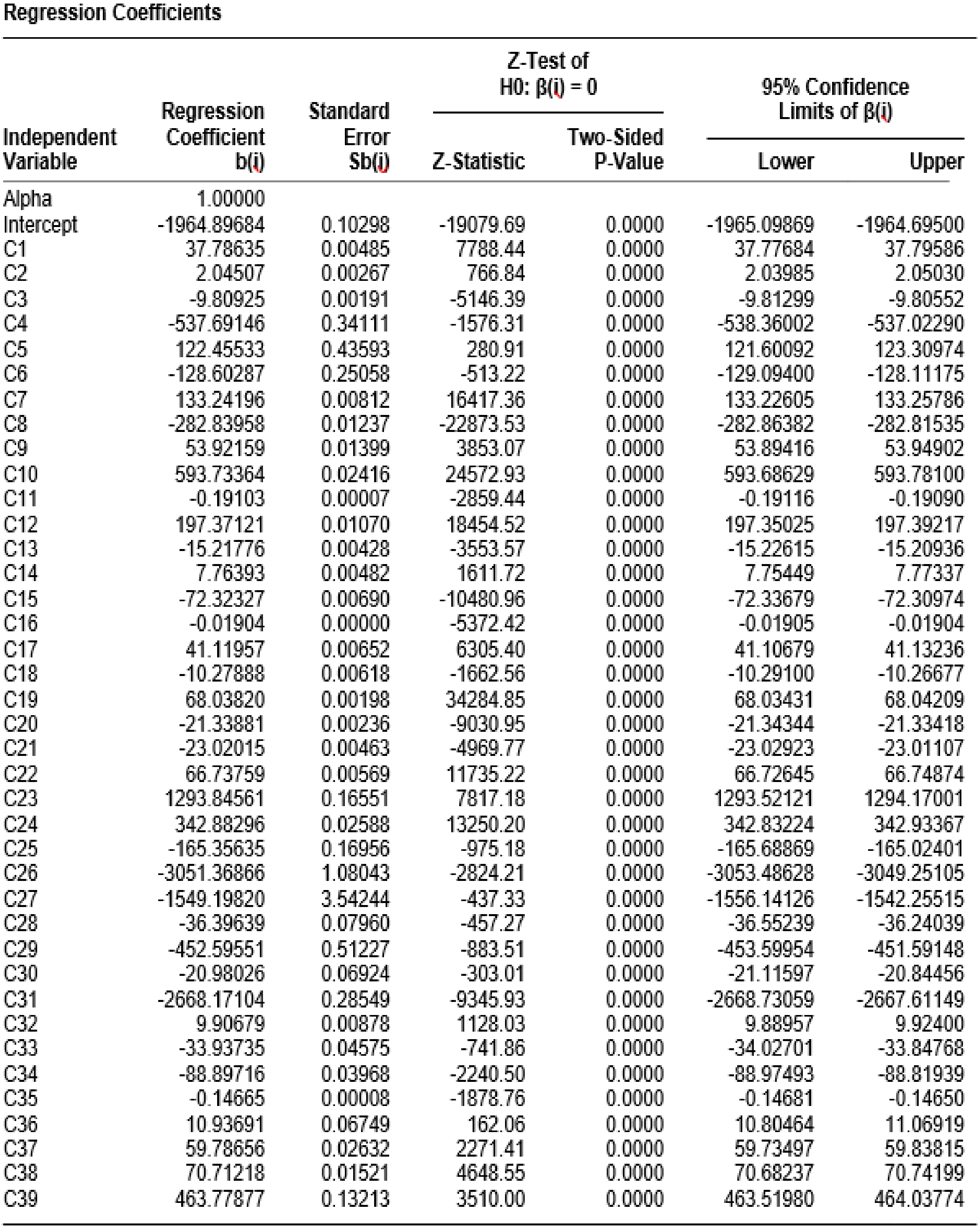


Regression Coefficients

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Z-Test of H0: β(i) = 0		95% Confidence Limits of β(i)	
			Z-Statistic	Two-Sided P-Value	Lower	Upper
Alpha	22.92689	0.00400	5737.53	0.000000	22.91906	22.93472
Intercept	105.59275	0.49312	214.13	0.000000	104.62624	106.55926
C1	-67.57674	0.02322	-2910.82	0.000000	-67.62224	-67.53124
C2	-9.95348	0.01276	-780.02	0.000000	-9.97849	-9.92847
C3	-12.20649	0.00914	-1335.30	0.000000	-12.22441	-12.18858
C4	1066.49488	1.63720	651.41	0.000000	1063.28602	1069.70374
C5	-7511.56876	2.58534	-2905.45	0.000000	-7516.63593	-7506.50159
C6	-389.92113	1.20180	-324.45	0.000000	-392.27662	-387.56565
C7	-30.09642	0.03882	-775.29	0.000000	-30.17250	-30.02033
C8	205.93464	0.05928	3473.91	0.000000	205.81845	206.05083
C9	137.47193	0.06700	2051.94	0.000000	137.34062	137.60324
C10	279.58357	0.11572	2416.14	0.000000	279.35678	279.81037
C11	0.06899	0.00032	215.85	0.000000	0.06837	0.06962
C12	54.88478	0.05118	1072.32	0.000000	54.78447	54.98510
C13	-27.44447	0.02050	-1338.93	0.000000	-27.48464	-27.40429
C14	-57.29652	0.02305	-2485.50	0.000000	-57.34170	-57.25134
C15	-68.16722	0.03302	-2064.51	0.000000	-68.23194	-68.10251
C16	-0.10595	0.00002	-6247.60	0.000000	-0.10598	-0.10591
C17	58.40590	0.03121	1871.52	0.000000	58.34473	58.46706
C18	202.85262	0.02958	6856.82	0.000000	202.79464	202.91061
C19	-4.39951	0.00949	-463.55	0.000000	-4.41811	-4.38091
C20	-67.02675	0.01136	-5898.87	0.000000	-67.04902	-67.00448
C21	-92.05289	0.02236	-4116.28	0.000000	-92.09672	-92.00906
C22	129.20147	0.02745	4707.28	0.000000	129.14768	129.25527
C23	213.02757	0.79251	268.80	0.000000	211.47428	214.58086
C24	186.35565	0.12403	1502.50	0.000000	186.11256	186.59875
C25	-356.76243	0.82684	-431.48	0.000000	-358.38301	-355.14185
C26	10558.08828	5.17154	2041.58	0.000000	10547.95224	10568.22431
C27	-48811.31718	16.99856	-2871.50	0.000000	-48844.63374	-48778.00061
C28	1338.92382	0.48511	2760.07	0.000000	1337.97303	1339.87461
C29	10345.07614	3.09309	3344.58	0.000000	10339.01380	10351.13848
C30	927.76617	0.33343	2782.48	0.000000	927.11266	928.41969
C31	-1127.29026	1.36648	-824.96	0.000000	-1129.96851	-1124.61202
C32	-27.76353	0.04204	-660.39	0.000000	-27.84593	-27.68113
C33	101.81815	0.21999	462.82	0.000000	101.38697	102.24933
C34	-124.14447	0.18990	-653.73	0.000000	-124.51667	-123.77227
C35	0.82404	0.00037	2199.54	0.000000	0.82330	0.82477
C36	601.45688	0.32301	1862.06	0.000000	600.82380	602.08996
C37	-197.08159	0.12602	-1563.85	0.000000	-197.32859	-196.83459
C38	-87.95467	0.07282	-1207.84	0.000000	-88.09739	-87.81194
C39	-1207.81870	0.63262	-1909.22	0.000000	-1209.05862	-1206.57878

The negative binomial regression is fitted in NCSS software. From table 2, we get the values of estimate, standard error, z value and p-value. From the table all the p-value are less than 0.05 so we can say that all variables are significant.

Table 3: Numerical Results Obtain by Performing Geometric Regression Model

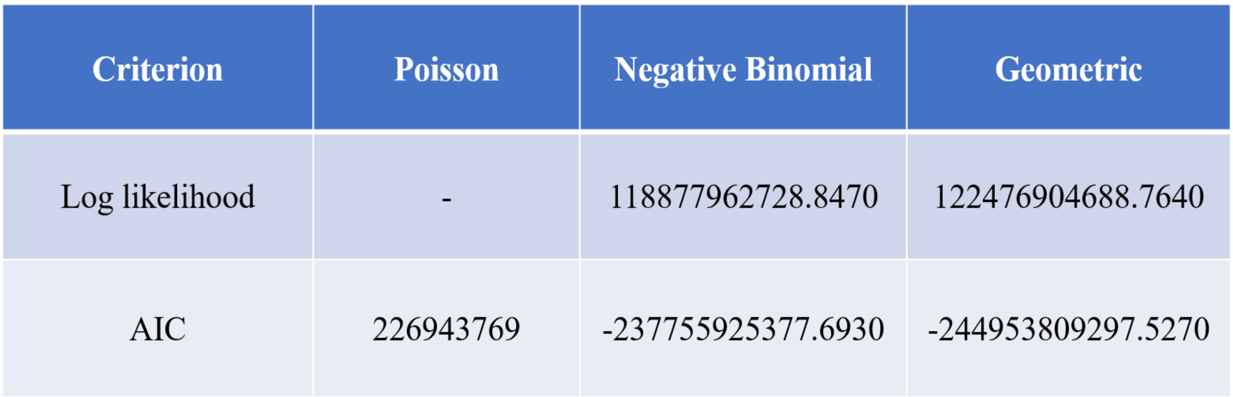


Regression Coefficients

			Z-Test of H0: β(i) = 0		95% Confidence Limits of β(i)	
Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Z-Statistic	Two-Sided P-Value	Lower	Upper
Alpha	1.00000					
Intercept	-1964.89684	0.10298	-19079.69	0.0000	-1965.09869	-1964.69500
C1	37.78635	0.00485	7788.44	0.0000	37.77684	37.79586
C2	2.04507	0.00267	766.84	0.0000	2.03985	2.05030
C3	-9.80925	0.00191	-5146.39	0.0000	-9.81299	-9.80552
C4	-537.69146	0.34111	-1576.31	0.0000	-538.36002	-537.02290
C5	122.45533	0.43593	280.91	0.0000	121.60092	123.30974
C6	-128.60287	0.25058	-513.22	0.0000	-129.09400	-128.11175
C7	133.24196	0.00812	16417.36	0.0000	133.22605	133.25786
C8	-282.83958	0.01237	-22873.53	0.0000	-282.86382	-282.81535
C9	53.92159	0.01399	3853.07	0.0000	53.89416	53.94902
C10	593.73364	0.02416	24572.93	0.0000	593.68629	593.78100
C11	-0.19103	0.00007	-2859.44	0.0000	-0.19116	-0.19090
C12	197.37121	0.01070	18454.52	0.0000	197.35025	197.39217
C13	-15.21776	0.00428	-3553.57	0.0000	-15.22615	-15.20936
C14	7.76393	0.00482	1611.72	0.0000	7.75449	7.77337
C15	-72.32327	0.00690	-10480.96	0.0000	-72.33679	-72.30974
C16	-0.01904	0.00000	-5372.42	0.0000	-0.01905	-0.01904
C17	41.11957	0.00652	6305.40	0.0000	41.10679	41.13236
C18	-10.27888	0.00618	-1662.56	0.0000	-10.29100	-10.26677
C19	68.03820	0.00198	34284.85	0.0000	68.03431	68.04209
C20	-21.33881	0.00236	-9030.95	0.0000	-21.34344	-21.33418
C21	-23.02015	0.00463	-4969.77	0.0000	-23.02923	-23.01107
C22	66.73759	0.00569	11735.22	0.0000	66.72645	66.74874
C23	1293.84561	0.16551	7817.18	0.0000	1293.52121	1294.17001
C24	342.88296	0.02588	13250.20	0.0000	342.83224	342.93367
C25	-165.35635	0.16956	-975.18	0.0000	-165.68869	-165.02401
C26	-3051.36866	1.08043	-2824.21	0.0000	-3053.48628	-3049.25105
C27	-1549.19820	3.54244	-437.33	0.0000	-1556.14126	-1542.25515
C28	-36.39639	0.07960	-457.27	0.0000	-36.55239	-36.24039
C29	-452.59551	0.51227	-883.51	0.0000	-453.59954	-451.59148
C30	-20.98026	0.06924	-303.01	0.0000	-21.11597	-20.84456
C31	-2668.17104	0.28549	-9345.93	0.0000	-2668.73059	-2667.61149
C32	9.90679	0.00878	1128.03	0.0000	9.88957	9.92400
C33	-33.93735	0.04575	-741.86	0.0000	-34.02701	-33.84768
C34	-88.89716	0.03968	-2240.50	0.0000	-88.97493	-88.81939
C35	-0.14665	0.00008	-1878.76	0.0000	-0.14681	-0.14650
C36	10.93691	0.06749	162.06	0.0000	10.80464	11.06919
C37	59.78656	0.02632	2271.41	0.0000	59.73497	59.83815
C38	70.71218	0.01521	4648.55	0.0000	70.68237	70.74199
C39	463.77877	0.13213	3510.00	0.0000	463.51980	464.03774

The geometric regression is fitted in NCSS software. From table 3, we get the values of estimate, standard error, z value and p-value. From the table all the p-value are less than 0.05 so we can say that all variables are significant.

Table 4: Numerical results of comparison of goodness of fit of Geometric, Poisson and Negative Binomial regression model



Criterion	Poisson	Negative Binomial	Geometric
Log likelihood	-	118877962728.8470	122476904688.7640
AIC	226943769	-237755925377.6930	-244953809297.5270

From table 4, we get the values of Log likelihood, AIC. Results of the three regression models are compared based on their respective, log likelihood and the AIC values as presented in above table.

Based on the model goodness-of-fit criterions, Poisson model appeared to outperform the other two models, as it has no log likelihood value and the high AIC value. NB model showed poor performance with lowest log likelihood and highest AIC value. Geometric model performs better than the NB model because it cannot accommodate the under-dispersion of the given data set. Geometric model can accommodate under-dispersion as well as over-dispersion.

Table 5: Numerical Results Obtain by Performing Poisson Regression Model for Under dispersed Variables

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	5.124e+00	1.598e-03	3206.11	<2e-16
n_tokens_title	4.158e-02	8.355e-05	497.68	<2e-16
num_keywords	1.763e-02	4.063e-05	433.99	<2e-16
n_unique_tokens	7.110e-01	3.621e-03	196.38	<2e-16
n_non_stop_words	9.706e-01	5.855e-03	165.78	<2e-16
n_non_stop_unique_tokens	-6.329e-01	3.498e-03	-180.93	<2e-16
num_hrefs	6.048e-02	1.238e-04	488.45	<2e-16
num_self_hrefs	-7.626e-02	1.537e-04	-496.02	<2e-16
num_imgs	7.292e-02	2.179e-04	334.74	<2e-16
num_videos	5.469e-01	3.903e-04	1401.10	<2e-16
kw_min_min	1.751e-01	1.444e-04	1212.94	<2e-16
kw_max_avg	-2.982e-03	8.339e-05	-35.77	<2e-16
kw_avg_avg	4.747e-02	2.500e-05	1898.52	<2e-16
is_weekend	1.050e+00	2.676e-03	392.43	<2e-16
global_subjectivity	3.210e-01	4.194e-04	765.40	<2e-16
global_sentiment_polarity	2.773e-01	2.732e-03	101.51	<2e-16
global_rate_negative_words	1.307e+01	6.255e-02	208.91	<2e-16
rate_positive_words	-1.923e-01	1.135e-03	-169.47	<2e-16
rate_negative_words	-2.026e+00	7.508e-03	-269.79	<2e-16
avg_positive_polarity	-1.887e-01	1.169e-03	-161.43	<2e-16
min_positive_polarity	-2.432e+00	4.813e-03	-505.39	<2e-16
max_positive_polarity	7.460e-03	1.490e-04	50.06	<2e-16
avg_negative_polarity	-1.970e-01	5.634e-04	-349.76	<2e-16
min_negative_polarity	2.461e-02	5.948e-04	41.38	<2e-16
title_subjectivity	-9.688e-02	1.129e-03	-85.78	<2e-16
title_sentiment_polarity	7.585e-02	4.176e-04	181.64	<2e-16
abs_title_subjectivity	6.982e-02	2.453e-04	284.62	<2e-16
abs_title_sentiment_polarity	5.647e-01	2.245e-03	251.54	<2e-16
global_rate_positive_words	-7.188e+00	1.894e-02	-379.48	<2e-16

We fit the Poisson regression for underdispersed variables in R software and we get the coefficients and p-values. All the p-values are less than 0.05 so we can say that all variables are significant.

Table 6: Numerical Results Obtain by Performing Negative Binomial Regression Model for over dispersed Variables

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	7.220e+00	4.271e-02	169.029	<2e-16
n_tokens_content	-2.633e-04	6.623e-04	-0.398	0.690969
average_token_length	-8.224e-04	4.935e-05	-16.664	<2e-16
kw_max_min	-1.519e-02	3.819e-03	-3.977	6.97e-05
kw_avg_min	6.205e-02	4.233e-03	14.657	<2e-16
kw_min_max	1.241e-02	6.789e-03	1.828	0.067624
kw_avg_max	6.543e-05	2.869e-06	22.810	<2e-16
kw_min_avg	-1.590e-02	6.491e-03	-2.449	0.014322
self_reference_min_shares	-3.961e-02	2.315e-03	-17.111	<2e-16
self_reference_max_shares	-7.823e-02	3.950e-03	-19.803	<2e-16
self_reference_avg_sharess	1.256e-01	5.208e-03	24.119	<2e-16
max_negative_polarity	-1.802e-04	5.200e-05	-3.464	0.000531

We fit the negative binomial regression model for over dispersed variables in R software and we get the coefficients and p-values.

Fitting of Regression Line Using Geometric Regression

Here we fit regression line using geometric regression model,

- **Geometric regression model** = np .exp (-1964.89684411877 + 37.7863501498931*6.487676 + 2.0450745525712*6.930885 - 9.80925231027686*12.41221 - 537.691461496132*0.494829 + 122.455332287826*0.704161 -128.602874205788*0.600249 + 133.241957043498*1.142021 - 282.839582128569*0 + 53.9215893081154*0 + 593.733642702667*0 - 0.191029690645754*471.5809 + 197.371210357818*0.577013 - 15.2177568254427*0 + 7.76393186112812*0.760199 -72.3232671555938*0 - 0.0190442272208242*892.9525 + 41.1195745017342*0.718003 - 10.2788840036801*16.21584 + 68.0382008693105*37.86738 - 21.3388105494879*0 -23.0201487016924*0 + 66.7375940711858*0 + 1293.84561211331*0 + 342.88295511799*1.09292 -165.356348392168*0.52501 - 3051.36866485426*0.028708 - 1549.19820471011*0.015401 - 36.3963909897244*1.360146 - 452.595511677002*0.333547 - 20.980264979223*0.843475 -2668.17103884373*0.123473 + 9.9067866567609*1.973298 - 33.9373459989945*1.557734 - 88.8971600287552*0.885116 - 0.146652698481876*215.1069 + 10.9369140599043*0.105433 + 59.7865570165658*0.904963 + 70.7121792932218*0.715185 + 463.778770263825*0)

  Predicted value of news shares using the above regression line is 4289.065248 and our actual value of news shares for the same values is 3600

Fitting of Regression Line for Under-dispersed and Over-dispersed Variables

- **Under dispersed Model** = np.exp (5.124042 + 0.041579*6.487675 + 0.017633*6.930885 + 0.06481*1.142021 - 0.076257*0 + 0.175144*0.577012 + 0.047470*37.867376 + 0.321046*1.092920 - 2.43226*0.123473 - 0.197037*1.557733 - 7.187985*0.028708 + 0.711037*0.494828 + 0.970568*0.704161 - 0.632923*0.6002493 - 0.002982*16.215835 +

0.277348*0.525010 + 13.066654*0.01540 - 0.192267*1.360146 - 2.025660*0.3335471 - 0.188666*0.843475 + 0.007460*1.973298 + 0.024610*0.88515 - 0.096881*0.105433 + 0.075847*0.904962 + 0.069824*0.715185)

Predicted value of the under dispersed Model using poisson regression is 1128.148341981

- **Over dispersed Model** = np.exp(7.220 - 0.0008224*471.5809 - 0.0159*0 + 0.06205*0.76019 + 0.00006543*892.952522 - 0.01590*0.718003 - 0.0001802*215.10689)

Predicted value of the over dispersed Model using Negative Binomial regression is 980.03573136 After fitting both regression lines individually and adding both the outputs we get the predicted value of news shares is 2108.184007 and our actual value of news shares for the same values is 3600.

7. CONCLUSIONS

In this project, for the count data, after fitting Poisson regression for under-dispersion variables, negative binomial for over-dispersion variables and geometric regression to resolve the both issues of under dispersion and over dispersion, it can be concluded that limitations of Poisson and Negative Binomial regression can be solved by Geometric regression. Using Poisson regression, only under dispersion problem can be solved. Poisson model appeared to outperform than other two models. Negative Binomial regression can only solve the problem of overdispersion. So, both Poisson regression and Negative Binomial regression can’t solve the under dispersion and over dispersion problem simultaneously.

Limitations of Poisson and Negative Binomial regression can be solved by Geometric regression. Geometric model can be used as an alternative modelling approach for both under-dispersed and over-dispersed count data. The fitting of the geometric regression model was found to be good. So, if the data is count then one can use the Geometric regression model to solve under-dispersion and over-dispersion problem simultaneously and using regression line equation, one can predict the count of shares of news. So, Geometric model may serve as an alternative model to Poisson and Negative Binomial models for modelling count data.

8. DISCUSSION

News is the most important part of our routine life. Nowadays everyone prefers reading news on online platforms. In this project the dataset is related to various variables which effect on the popularity of news shared. Data for this project contained some under dispersed variables and over dispersed variables. Poisson regression is used for under-dispersion variables, negative binomial for over-dispersion variables and geometric regression was performed to resolve the both issues of under dispersion and over dispersion.

From this project, we can suggest so many things so that news will share more and more and there will be chance to become news popular. From various graphs it is observed that news related to world and lifestyle are mostly shared. Most of the news are shared on Tuesday, Wednesday, and Thursday. From scatter plots it is observed that the number of words in the article should be less than 1500 words. The lesser the better. Article title shouldn't be too long or too short. 6 – 17 words is the ideal number of words to have for titles. Articles should have good number of images. Between 1 – 40 images are great. Number of videos should be between 0 - 25 videos. The number of keywords in the metadata really influences the shares to a margin. The higher the value the better the shares chances. A value upward of 5 is recommend. Using these suggestions one can write news properly so that it gets popular.

In this project, after fitting Poisson regression for under-dispersion variables, negative binomial for over-dispersion variables and geometric regression was performed to resolve the both issues of under dispersion and over dispersion, it can be concluded that limitations of Poisson and Negative Binomial regression can be solved by Geometric regression. Geometric model can be used as an alternative modelling approach for both under-dispersed and over-dispersed count data. So, if the data is count then one can use the Geometric regression model to solve under-dispersion and over-dispersion problem simultaneously and using regression line equation, we can predict the count of shares of news.

9. FUTURE WORK

When we fit the geometric model, we don’t get proper answers for the predicted values of the news shares. There is difference between actual value and predicted value. So, we need more work for exact prediction.

In this project we try to give probability estimates other than MLE or ME. Poisson regression line under-dispersed variables and negative binomial regression line for over-dispersed variables and we combine the results, we get a significant difference in the actual and predicted values of the news shares. And we don’t know whether combining the results of two different models is proper method or not. So, we need to work more on our data, model fitting and try to sort the issues we are facing currently.

REFERENCES

- Al-balushi, z. M. D., & Islam, m. M. (2020). Geometric regression for modelling count data on the time-to-first antenatal care visit. *Journal of statistics: advances in theory and applications*, *23*(1), 35-57.
- Bandari, R., Asur, S., & Huberman, B. (2012). The pulse of news in social media: Forecasting popularity. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 6, No. 1, pp. 26-33).
- Fernandes, K., Vinagre, P., & Cortez, P. (2015). A proactive intelligent decision support system for predicting the popularity of online news. In *Progress in Artificial Intelligence: 17th Portuguese Conference on Artificial Intelligence, EPIA 2015, Coimbra, Portugal, September 8-11, 2015. Proceedings 17* (pp. 535-546). Springer International Publishing.
- Harris, T., Yang, Z., & Hardin, J. W. (2012). Modeling under dispersed count data with generalized Poisson regression. *The Stata Journal*, *12*(4), 736-747.
- Lindén, A., & Mäntyniemi, S. (2011). Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology*, *92*(7), 1414-1421.
- Petrovic, S., Osborne, M., & Lavrenko, V. (2011). Rt to win! predicting message propagation in twitter. In *Proceedings of the international AAAI conference on web and social media* (Vol. 5, No. 1, pp. 586-589).
- Szabo, G., & Huberman, B. A. (2010). Predicting the popularity of online content. *Communications of the ACM*, *53*(8), 80-88.
- Tatar, A., Antoniadis, P., Amorim, M. D. D., & Fdida, S. (2014). From popularity prediction to ranking online news. *Social Network Analysis and Mining*, *4*, 1-12.
- Yalcin, I., & Amemiya, Y. (2001). Nonlinear factor analysis as a statistical method. *Statistical science*, 275-294.