Evaluation of Google Vision API for Object Detection in General Subject Images

Leah Schultz
lschult@tarleton.edu
Department of Marketing and Computer Information Systems
Tarleton State University
Stephenville, Tx 76401 USA

Mark Adams
madams@tarleton.edu
Department of Accounting, Finance, and Economics
Tarleton State University
Stephenville, Tx 76401 USA

Abstract

This paper examines and compares Google's Vision application programming interface against a data set of keywords provided by human participants. Students and Google Vision provided labels for 40 images that were divided into four category types: landscape, portrait, news, and cityscape. The terms provided by student participants and Google were then compared. Results indicate that Google provides, on average, more terms per image than the human participants in the study. This held true across all image types. Results also indicate that there were low levels of agreement on terms between Google and participants. Reasons for disparities between number of terms provided and the low level of agreement are discussed as well as implications for general subject image retrieval and future research.

Keywords: computer vision, image retrieval, Google Vision

1. INTRODUCTION

To celebrate International Cat Day, Facebook conducted a study to determine which of its users were cat people and which of its users were dog people based on the images that they shared on their timelines. Facebook used a technology they are currently developing using object recognition technology. Unlike facial recognition software which has been used widely across many fields, object recognition software has lagged behind in development. Instead, many specialized recognition software applications have been developed in narrowly defined fields such as medicine and agriculture (Adamic, Burke, Herdagdelen & Neumann, 2016).

Imagine, however, if image recognition software became more efficient and could be applied to many different fields. For instance, given the scenario above, consider the use of this kind of data in marketing. In addition to being useful to the billions of dollars a year pet food industry, imagine the data that would become available from a user's Instagram account that mainly consists of image data. How could businesses use the data that indicates that cat people, as defined in Facebook's study are more likely to watch Dr. Who versus their dog counterparts who prefer programming such as The Voice? The ability to mine data from image sources could be useful to businesses in many ways from analyzing trends to assessing brand sentiment.

In order to do this effectively, businesses need access to tools that can accurately and efficiently analyze images using advances in computer vision. Google recently released its Google Vision to help customers do exactly that. This paper will compare the results of Google Vision’s label functionality to labels provided by human subjects.

2. BACKGROUND

Image Analysis

Determining what an image is about or assigning search terms for retrieval is a field that has a long history in library science and a newer, rich field of literature in computer vision. The two approaches to the retrieval differ in the use of a human component to provide search terms to identify images versus the artificial intelligence algorithms of computer science to automatically detect objects in images. The first approach is steeped in traditional library science with controlled vocabularies and structured search terms that have fairly accurate results but are time consuming and not always effective in retrieval situations (Shatford, 1984, Jörgensen , 1999, Choi & Rasmussen, 2003; Jörgensen ,2004; Jörgensen and Jörgensen; 2005). A newer approach to human supplied search terms is the concept of crowdsourcing which has been used in specialized fields such as medicine and agriculture but also in general subject matter image databases (Xian-Hong et al, 2014; Borgo, Micallef, Bach, McGee & Lee, 2018) In fact, many general use applications such as stock photography databases or even social media sites such as Instagram depend on the creator of the image to supply the context. Because there have been numerous studies on sentiment analysis of text, this many times is a fall back solution to determining the content of the image through the text that is associated with it.

Automatic analysis of image content is not a new field and early studies were limited to the physical characteristics of an image such as color or texture. These fields grew quickly in sophistication as they were applied to many technical fields that sought improvements in accuracy and efficiency for image analysis (Goodrum, Rorvig, Jeong, and Suresh 2001). Many medical fields have been transformed by image analysis. Interpretation of medical imagery from x-rays to CAT scans have been improved through use of computer vision (Gonçales, Guilherme, Pedronette, 2018; Ximei et al, 2017). Fields such as agriculture have deployed computer vision applications to identify pests and to recognize disease. Facial recognition which is a subset of computer vision studies have been implemented in many fields from criminal justice, security, and social media (Blanco-Gonzalo, Lunerti, Sanchez-Reillo, & Guest , 2018; Ricanek & Boehnen, 2012; Wang et al, 2017). Interest in computer vision continues to grow as technology in areas such as autonomous driving and robotics continues to increase.

Recent research has begun to apply these technologies to more generalized subject matter in image collections (Li, Purushotham, Chen, Ren, & Kuo, 2017). A growing field of researchers are creating tools to predict sentiment in images (Seo & Kang, 2016; Soleymani et al, 2017; Bai, Chen, Huang, Kpalma, & Chen, 2018). Other researchers are studying automatic trademark retrieval to apply in business fields (Anuar, Setchi, & Lai, 2013).

Some approaches to image recognition incorporate both a human and a technology component. Researchers use human supplied terms or behavior to inform the algorithms used in automated systems. For example, Google uses crowdsourcing volunteers to provide photos as well as captioning for their photos and then uses this information to inform the artificial intelligence application to group like photos and assign similar search terms (Google, 2018a). Researchers in Japan are taking a different approach by scanning users’ brains while viewing images. These scans are then analyzed using an algorithm to develop descriptive phrases of the images based on the brainwaves of participants.

Google Vision API

The tool used in this study was developed by Google and released for public use in response to Microsoft’s Cognitive Services API for image, speech, and semantic analysis. Google Vision is a cloud based application that allows developers to use the features of the product in their own applications through an application programming interface. Smaller sets of images can be processed individually on the Vision API website. The price of the product ranges from free for a limited number of images up to enterprise pricing packages. The packages consist of different functionality associated with image recognition. Label detection provides keywords or labels to describe the images provided. In addition to labels, Google has specific technology to detect logos, landmarks, text, and face detection in images. The face detection functionality does not include facial recognition but provides information on emotional state and headgear. Additional functionality of the product includes identification of explicit content which can be used to monitor and block unwanted adult content on sites. Google Vision also provides attribute analysis such as dominant colors and also provides cropping advice. Finally, Google Vision uses the power of Google's search engine to locate similar images on the internt. (Google, 2018b)

3. METHODOLOGY

Data from a previous study (Schultz, 2009) was used to compare with the results of image recognition software available in Beta testing from Google. From the previous study, 61 students were instructed to provide terms to describe the images. Participants were further instructed to provide as many terms as they felt necessary to describe the image. Students viewed 40 images selected from government collections, stock photography collections, a news agency, and personal images belonging to the researcher. Students in the study were all undergraduate students from a regional public institution. Students were, on average, 22.4 years of age (*SD*=4.9) and represented various ethnicities and academic backgrounds. Students were almost equally divided by gender with 29 female and 31 male participants. One participant chose not to disclose their gender.

The same 40 images were submitted to Google Vision API and the results for the landmarks and labels data were retrieved. Data from other parts of the results were not included. Terms were saved from the tool and placed into the same database as the terms from the students in the study.

Images used in the study were not focused on one area of content and did not require specialized knowledge of the participant. The study utilized images that were divided into four distinct categories: news images, landscapes, portraits, and cityscapes. The images contained a variety of objects, people, and locations.

The final data set includes 2,472 observations including 2,432 observations from the prior experiment and 40 observations taken from Google's tool. The 2,432 observations were the results of the 61 students responding to the 40 images. The response variable is the number of keywords reported by each participant/Google for each image. 8 observations from human participants were excluded because there were no keywords supplied. The 40 images are grouped into the 4 types, consisting of landscape, portrait, cityscape and news. The two-way ANOVA model is constructed as follows:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha_i\beta_j + \epsilon_{ijk,}$$

where ɑ indicates participant type (Google or human) and β indicates image type. The test is configured to test four pairwise comparisons specifically to test human versus Google for each image type.

4. FINDINGS

Initial analysis of the data found that participants provided on average 170.97 terms while Google provided 682 terms. Per image, that results in participants providing 4.27 terms per image while Google generated 17.05 terms per image.

Tests of normality show that the sample data is not normalized, with the Shapiro-Wilks test rejecting the null hypothesis of normality at 0.000. Further graphical evidence of normality issues can be seen in Appendix A. The data also exhibits heteroscedasticity, with the Breusch-Pagan test rejecting the null hypothesis of equal variance at 0.000. Transformations of the response variable to correct for unequal variance were not successful. Finally, since there is not a time-series data set, there is no statistical test to evaluate the independence assumption, but the data taken from the experiment is independent, the order of images given to participants was randomized to remove order effects, and each participant provided keywords for each image only once.

Results from ANOVA, available in Appendix B, yield evidence that both main effects of participant type (human or Google) and image type are significant at 0.05, as is the interaction of the types ($p<0.01$ for all related F-tests), so further reporting will be based on treatment means. A priori pairwise comparisons of participant types within each of the four image types were prepared and 95% simultaneous confidence intervals were obtained using the Bonferroni method. Confidence intervals for each image type are listed in Table 1. All confidence intervals are simultaneously significant at 95% and indicate that Google provided more keywords per image than the human participants for all image types. The differences were very large for cityscape, landscape, and portrait (point estimates of -16.45, -20.87 and -

10.19 respectively), but were noticeably smaller for news type images, with a point estimate of -3.4. Table 1 shows the average difference in number of terms for each image type. For example, for cityscape image, Google produced between 14 and 18 more terms on average than human participants. This suggests that further research may identify an area where Google may be able to improve its image recognition algorithm.

	(1)	(2)
Image Type	95% LCL	95% UCL
Cityscape	-18.164	-14.734
Landscape	-22.589	-19.159
News	-5.253	-1.822
Portrait	-11.901	-8.471

95% simultaneous confidence intervals using the Bonferroni method. For each image type, C.I. values are calculated for human participant submission - Google submission.

Table 1

An initial analysis of overlap in terms provided between Google and human participants shows a pattern of little agreement on the terms needed to describe the provided images. Only 9 of the terms provided by Google matched terms provided by participants at a level of 30% or more, with 70% being the highest level of agreement with 42 of the 60 students providing the term "family" for one of the images. On the lower end of the spectrum, 68.32% (466 out of 682) of the terms provided by Google were not matched at all by a human participant.

5. DISCUSSION

This initial comparison of Google Vision API labels against those provided by human participants provides a glimpse into strengths and weaknesses of both approaches as well as problems and opportunities in the field of image recognition.

More Terms for Google

As outlined in the findings, it appears that Google provided quite a few more terms on average than its human counterpart. At almost 4 times the terms per image, Google appears to have either named more items in images or just worked harder to assign meaning than the students participating in the project. This however, may be impacted by the quality or level of terms that were provided, something that was not directly addressed in this study. For example, Google would provide terms at all levels on a hierarchy of identification. Given a picture of a waterfall in a forest, students were more likely to identify the trees in the picture whereas Google not only identified the trees but also more generic terms such as vegetation and very specific terms such as leaf. In addition to the specificity of parts, Google also provided the scientific names of some of the plants in the images. The way people picture the whole instead of individual components may differ slightly than how the computer analyzes an image. For example, a portrait of an image might result in a few students providing the term face but it is unlikely they would provide terms for all components of the face such as the eyes, eyelashes, nose, etc.

Whether this is a strength of the system for its thoroughness or a weakness which provides terms that do not add value to the identification of the image, remains to be seen and would probably be specific to the context of the user retrieving or analyzing the image. On the surface, it appears that Google is more thorough in listing items whereas the student participants were more likely to provide terms that would be useful for retrieval purposes. Students were not exhaustive in their assignment which would be an advantage of the automated system. However, there is some decision making on students' behalf to pick the most important or added judgment as to the "best" words. Google also adds a level of

What's interesting to note is that the type of image being analyzed appears to have an effect of how thorough Google is in providing labels compared to human subject. On all types of images, Google supplied more terms but it was a noticeably smaller gap on images that came from news stories. These images which many times included more objects and more action than other images tended to have fewer labels applied. For example, in one image where a body is being removed from a crime scene, Google only provides the terms vehicle, car, profession, and laborer. The ambulance, the body, the police tape are all overlooked or not identifiable. Perhaps these images held the human eye a little longer for consideration as they told a story as opposed to a landscape or a portrait.

Performance

Surprisingly, the initial analysis of matches between the keywords provided by Google and the students participating in the process seems low. With almost 70% of the terms that Google provided not matching a terms provided by a student, it appears that the results are not very useful. This, however, needs additional analysis.

In addition to the discussion of level of details described in the last section, one also needs to take into consideration the wide variation of terms that differ very little in meaning. The terms provided by the student were not controlled for spelling or variation and the initial analysis at this stage was looking for exact matches. For example, one label provided by Google was "rock" which would not have matched up with a term had a student typed "rocks" instead. Additional study needs to be done to determine if controlling for some of these issues would improve the rate of matching significantly.

Another hurdle to higher levels of matching is the way that Google phrases some of the labels it provides. For example, on one image of a farmhouse in front of a gathering storm, both Google and almost 30% of the students provided the term "storm". Google also provided the term "meteorological phenomenon" and landscapes contained "coastal and oceanic landforms". Again, this will obviously affect the rate of matching but these terms could potentially be useful in a certain context and may not indicate poor performance.

As for accuracy, there were problems with both Google's API and the human subjects in the study. An image of a canyon in Arizona received labels of leg, human body, and muscle as Google had difficulty distinguishing the patterns on the canyon walls. An image of a man with shoulder length hair was improperly identified as a girl which is not a mistake that would likely be made by a human due to other indicators of gender that are more difficult to escape human notice.

Where Google did excel in identification over its human counterparts was in identifying locations and landmarks. Only one person was able to identify the skyline of Chicago which Google did with ease. Even more obscure locations such as Antelope Canyon and the construction site of the new World Trade Center were easily identified by Google. Neither of these images were identified correctly by the students in the study.

Finally, the main purpose of the API is to identify objects in the image. Google does provide some labels that do not correspond to objects necessarily. From extrapolating details in the image, Google provides more abstract terms such as "vacation" and "learning", and, on a few images, provides even more abstract ideas such as emotion in the form of the terms "fun", "joy", and "sorrow". Looking at the corresponding terms in the students' responses, there are many more occurrences of abstract concepts and story-telling. Again, this is neither good nor bad but a marked difference in the approach in identification. The story presented by the student may be erroneous and have little value in analyzing the image. However, the tone or emotion of an image could be something that would be of interest to users, particularly when trying to analyze user sentiment.

6. CONCLUSIONS

As improvements are made in general use image analysis, more and more uses for this data will be realized. From the marketer trying to analyze who is using their products and what context they are using the product could be helpful to reach new audience, improve marketing efforts to existing customers, or predict trends in the market. Social media sites that depend more on image data than text data such as Instagram and Snapchat become minefields of data for the social sciences. Historical images and collections of art become searchable in ways that have not been available to scholars before. With the amount of image data being uploaded to the internet each day, it's difficult to predict how this rich source of information will be used.

This study examines one aspect of the functionality and the results of the one API against a small group of human subjects. Further research needs to be conducted with larger groups to determine if there are regional or cultural differences that might differ from the group studied. In addition to that information, other systems developed by other companies could be compared using the same results from the humans providing terms. In addition to studying other algorithms used in image detection, additional methods need to be developed to be able to compare systems, both automated and human directed, controlling for language variations and levels of confidence or importance assigned by various identifiers.

7. REFERENCES

Adamic, L., Burke, M., Herdağdelen, A, & Neumann, D. (2016) Cat People, Dog People in *Facebook Research*. Retrieved June 6, 2018 from http://research.fb.com/cat-people-dog-people/

Anuar, F., Setchi, R., & Lai, Y. (2013). Trademark image retrieval using an integrated shape descriptor. *Expert Systems With Applications*, 40(1), 105-121.

Bai, C., Chen, J., Huang, L., Kpalma, K., & Chen, S. (2018). Saliency-based multi-feature modeling for semantic image retrieval. *Journal Of Visual Communication & Image Representation*, 50199-204.

Blanco-Gonzalo, R., Lunerti, C., Sanchez-Reillo, R., & Guest, R. M. (2018). Biometrics: Accessibility challenge or opportunity?. *Plos ONE*, 13(3), 1-20.

Borgo, R., Micallef, L., Bach, B., McGee, F., & Lee, B. (2018). Information Visualization Evaluation Using Crowdsourcing. *Computer Graphics Forum*, 37(3), 573-595.

Choi, Y. & Rasmussen, E. (2003). Searching for images: The analysis of users' queries for image retrieval in American history. Journal of the American Society for Information Science and Technology, 54(6), 498-511.

Gonçalves, F. F., Guilherme, I. R., & Pedronette, D. G. (2018). Semantic Guided Interactive Image Retrieval for plant identification. Expert Systems With Applications, 9112-26.

Google (2018a). Crowdsource. Retrieved June 6, 2018 from http://crowdsource.google.com/

Google (2018b) Cloud Vision API. Retrieved June 6, 2018 from http://cloud.google.com/vision/

Goodrum, A.A., Rorvig, M.E., Jeong, K., & Suresh, C. (2001). An open source agenda for research linking text and image content features. Journal of the American Society for Information Science and Technology, 52(11), 948-953.

Jörgensen, C. (1999). Access to pictorial material: A review of current research and future prospects. Computers and the Humanities, 33, 293-318.

Jörgensen, C. (2004). The visual indexing vocabulary: Developing a thesaurus for indexing image across diverse domains. Proceedings of the 67th ASIS&T Annual Meeting, 41, 287-293.

Jörgensen, C. & Joregensen, P. (2005). Image querying by image professionals. Journal of the American Society for Information Science and Technology, 56(12), 1346-1359.

Li, S., Purushotham, S., Chen, Ren, Y., & Kuo, C. J. (2017). Measuring and Predicting Tag Importance for Image Retrieval. IEEE Transactions On Pattern Analysis & Machine Intelligence, 39(12), 2423-2436.

Ricanek, K., & Boehnen, C. (2012). Facial Analytics: From Big Data to Law Enforcement. *Computer*, 45(9), 95-97.

Schultz, L.A. (2009). Image manipulation and user-supplied index terms (Doctoral dissertation). Retrieved from http:// digital.library.unt.edu/ark:/67531/metadc9828/.

Seo, S., & Kang, D. (2016). Study on predicting sentiment from images using categorical and sentimental keyword-based image retrieval. *Journal Of Supercomputing*, 72(9), 3478-3488.

Shatford, S. (1984). Describing a picture: A thousand words are seldom cost effective. *Cataloging and Classification Quarterly*, *4*(4), 13-30.

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S., & Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image & Vision Computing*, 653-14.

Wang, X., Guo, G., Merler, M., C. F. Codella, N., MV, R., Smith, J. R., & Kambhamettu, C. (2017). Leveraging multiple cues for recognizing family photos. *Image & Vision Computing*, 5861-75.

Xian-Hong, X., Xiao-Yu, H., Xiao-Ling, Z., Chun-Fang, C., Jian-Yong, Y., & Lei, L. (2014). Many Can Work Better than the Best: Diagnosing with Medical Images via Crowdsourcing. Entropy, 16(7), 3866-3877.

Ximei, H., Jianjie, B., Nan, Z., Xiaoling, D., Fei, L., & Fadong, H. (2017). Application of Computer Vision Technology in Agriculture. *Agricultural Science & Technology,* 18(11), 2158-2162.

ISSN: 2167-1508
v11 n 4814

APPENDIX A – GRAPHICAL EVIDENCE OF ASSUMPTIONS

Figure 1- Box plot of response variable

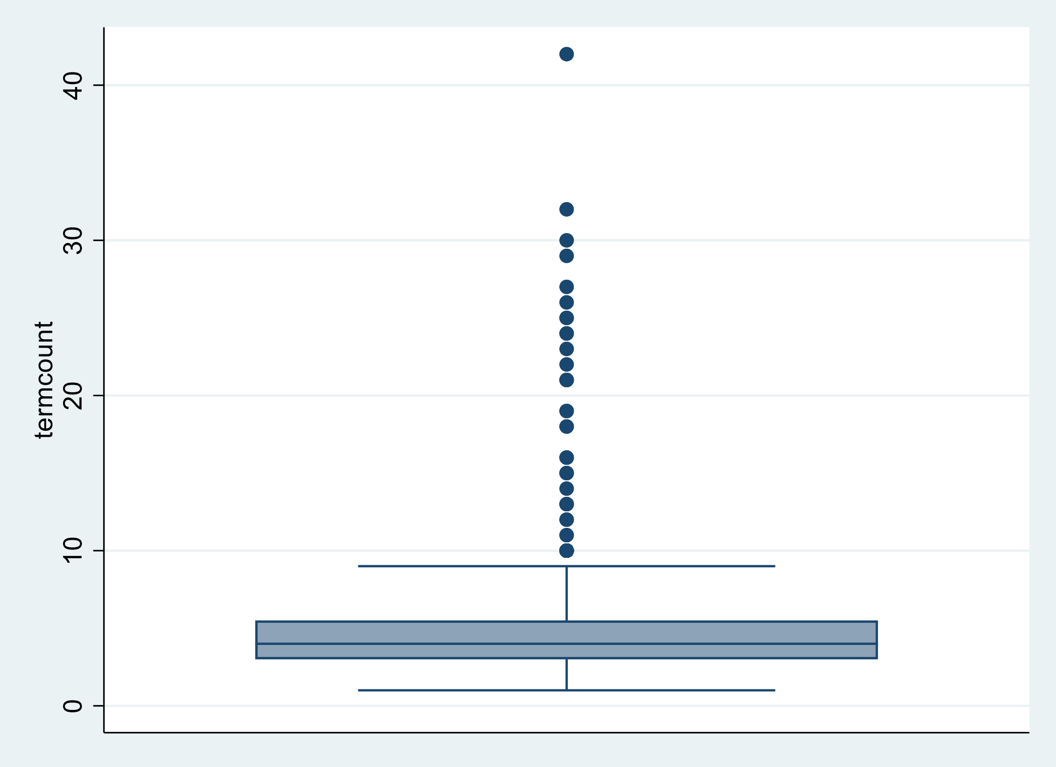


Figure 2 - Normal Probability Plot of predicted values

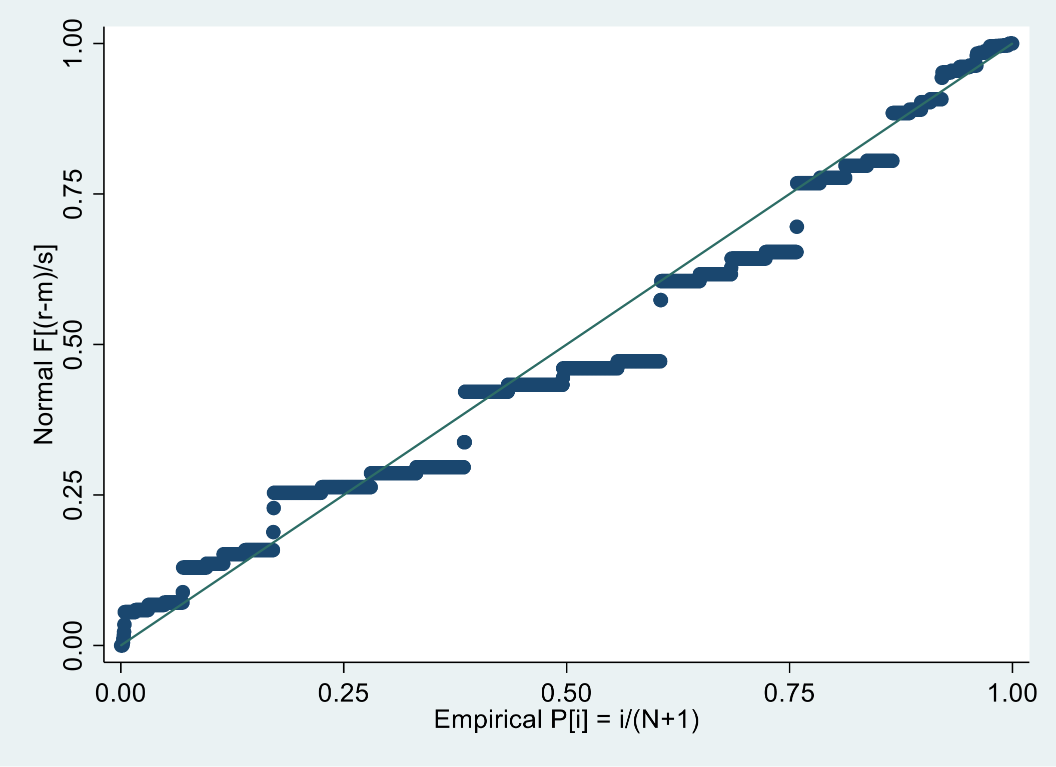
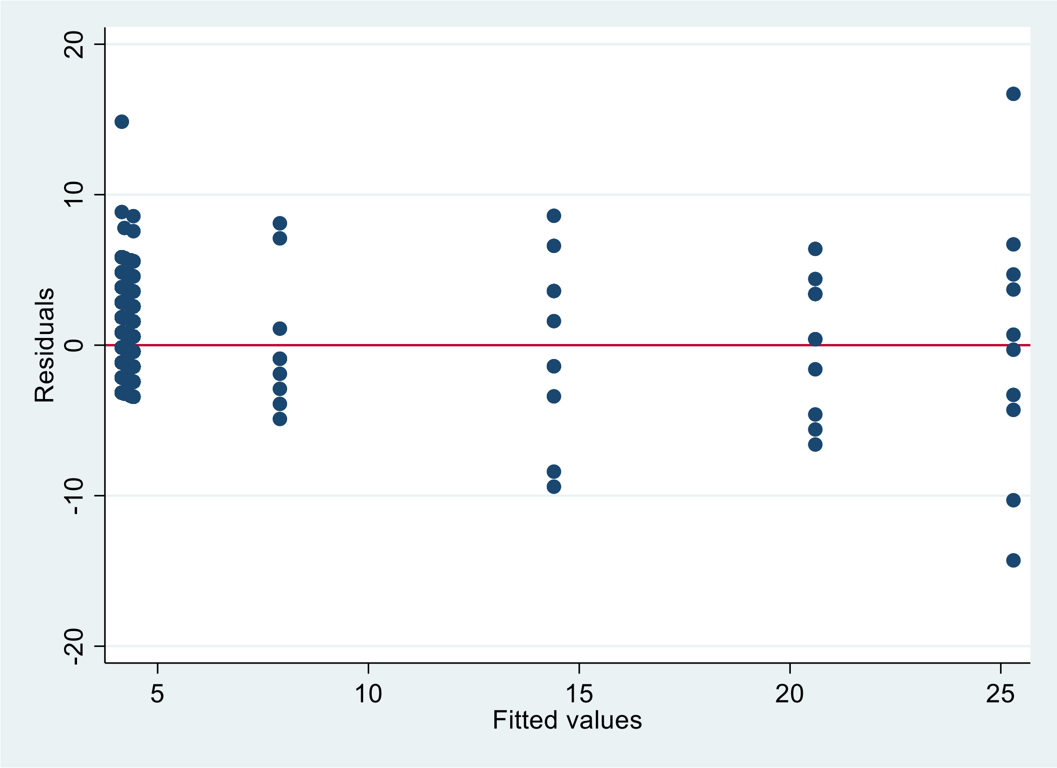


Figure 3 - Residual vs. Predictor plot



Appendix B – Statistical test results

ANOVA

```
                         Number of obs =      2,472    R-squared     =  0.4166
                         Root MSE      =    2.15238    Adj R-squared =  0.4150

                  Source | Partial SS         df         MS        F    Prob>F
        -----------------+----------------------------------------------------
                   Model |  8152.8929          7    1164.699    251.41  0.0000
                         |
               imagetype |  1690.1265          3   563.37551    121.61  0.0000
                  isgoog |  6409.0005          1   6409.0005   1383.42  0.0000
        imagetype#isgoog |  1683.5246          3   561.17485    121.13  0.0000
                         |
                Residual |  11415.039      2,464   4.6327268
        -----------------+----------------------------------------------------
                   Total |  19567.932      2,471   7.9190334
```

Bootstrap values for F-score of model and all model variables, and residual/error partial SS

```
------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     F_model |   251.4068   60.19838     4.18   0.000     133.4201    369.3935
         F_1 |   121.6078    42.2913     2.88   0.004     38.71833    204.4972
         F_2 |   1383.419   297.2092     4.65   0.000     800.8992    1965.938
         F_3 |   121.1327    41.7131     2.90   0.004     39.37655    202.8889
         rss |   11415.04   562.6597    20.29   0.000     10312.25    12517.83
------------------------------------------------------------------------------
```

Treatment Mean and Standard Error section

```
------------------------------------------------------------------------------------
                    |            Delta-method
                    |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------------+---------------------------------------------------------------
   imagetype#isgoog |
       cityscape#0  |   4.151067   .0872187    47.59   0.000     3.980038    4.322097
       cityscape#1  |       20.6   .6806414    30.27   0.000     19.26531    21.93469
       landscape#0  |   4.425987   .0872904    50.70   0.000     4.254817    4.597157
       landscape#1  |       25.3   .6806414    37.17   0.000     23.96531    26.63469
            news#0  |   4.362438   .0873623    49.94   0.000     4.191127    4.533749
            news#1  |        7.9   .6806414    11.61   0.000     6.565312    9.234688
        portrait#0  |   4.213816   .0872904    48.27   0.000     4.042646    4.384986
        portrait#1  |       14.4   .6806414    21.16   0.000     13.06531    15.73469
------------------------------------------------------------------------------------
```

Table 1 - Planned Contrasts

	µ A1 B1	**µ A1 B2**	**µ A2 B1**	**µ A2 B2**	**µ A3 B1**	**µ A3 B2**	**µ A4 B1**	**µ A4 B2**		**L̂**	**B**	**Std err**	**MSE**	**LCL**	**UCL**
X̅	4.151	20.6	4.426	25.3	4.362	7.9	4.214	14.4			0.993		4.632		
Sample sizes	609	10	608	10	607	10	608	10			2.499				
L1	1	-1								-16.44	2.499	0.686		-18.164	-14.734
L2			1	-1						-20.87	2.499	0.686		-22.589	-19.159
L3					1	-1				-3.537	2.499	0.686		-5.253	-1.822
L4							1	-1		-10.18	2.499	0.686		-11.901	-8.471