Snappy title goes here

Charities, funding, and Twitter

|  |  |  |
| --- | --- | --- |
| Natalie Polack University of Stirling npp00002@students.stir.ac.uk | Tom Wallace University of Stirling tom.wallace@stir.ac.uk | Vikki Richardson University of Stirling vikkirichardson@hotmail.co.uk |

ABSTRACT

T

KEYWORDS

A

1 INTRODUCTION

This research paper explores factors that predict charity use of Twitter. Twitter is a social networking platform that is commonly used by charities but remains understudied academically (Obar, Zube, & Lampe, 2012). Charities have many uses for Twitter, including advertising charitable campaigns, interacting with their users or supporters, and even communicating directly with other charitable organizations. Twitter has a relatively low cost of use, only requiring staff time. This may present an opportunity for smaller charities, who are often limited by costs, but larger charities are more likely to have the capacity to dedicate staff to social media and so are still, likely, at an advantage. Our study attempts to predict various aspects of charity Twitter usage (such as overall use, popularity, and networking activity) based on publically available data on charities features (such as funding source, overall income, who the charity helps, and number of staff). This study is important because there is a lack of published research on the factors which relate to successful use of social media by charities. Twitter has huge potential for charitable organisations and knowing which factors contribute to success or failure on the platform could help implement measures that lead to more of the former and less of the latter.

The next part of this paper give a brief summation the research questions and data management tasks. Section 2 then describes pervious literature and research in this area. Sections 3 and 4 detail the data processing and analysis methods used before the results are discussed in section 5. The paper concludes with a summary of the implications of the findings in section 6.

2.2 Research questions

**RQ1**: Is funding source related to use of Twitter?

**RQ2**: Are charities which seek to help the public more popular on Twitter?

**RQ3**: Does number of staff, rather than size, determine a charity’s active use of Twitter?

2.3  Data sources and management tasks

Three data sources were required to undertake this research. To obtain data on the outcome, Twitter use by charities, we downloaded data directly through the Twitter API (2018). For data on charity features, we scraped the UK Charity Commission website (2018) which holds a record for every registered charity in England and Wales (Scotland and Northern Ireland are covered by separate regulators). Given these two data sources, we could sample an arbitrary number of charities, right up to the full population of ~168,000 that are registered. However, the Charity Commission website does not record detailed financial information for charities which is important for RQ1. Therefore, we make use of data collected by Alcock and Mohan of the Third Sector Research Centre (2017). This dataset is based on surveying a random sample of registered charities and contains detailed financial information, such as funding source. As this dataset is the smallest, it forms our sample and we use it as a ‘mask’ to collect data from the Charity Commission and Twitter on this subset of charities. The Alcock and Mohan data gives us a sample of 12,150. Each of these data sources needed to be interacted with in different ways and presented different data management challenges which are detailed in section 3.

Because the Alcock and Mohan data provided the sample, it was sourced and managed first. The Charity Commission data was then scraped which provided the websites which were used to collect Twitter handles. Finally, these handles were used to collect Twitter data through the API. The data sets were then recombined to form a final analysis dataset which was analysed in parallel as shown in the code tree below.

[Code tree diagram]

2 SCENARIO

2.1 Literature and related work

Twitter use by charities is common. The Lloyds Digital Index found that 44% of the, more than one hundred thousand, charities they surveyed in 2016 were using social media in some form (Lloyds Bank, 2016). Of these platforms, Twitter is by far the most used platform for charitable organisations (Guo & Saxton, 2014). But what do charities use Twitter for? A common use case is broadcasting one-to-many messages to followers to convey information or solicit donations (Phethean, Tiropanis, & Harris, 2015; Waters & Jamal, 2011). Another use of Twitter is networking activity between charities, either because they share a common purpose or for support. Infrastructure organisations, which are charities established to help support other charities, can use Twitter to share links to information, resources, and training (Dayson & Sanderson, 2014).

Overall, the literature, which has been published so far, seems to concur that social media use helps make a charity successful (Lloyds Bank, 2016; McCabe & Phillimore, 2012). Therefore, it is important to understand what factors make a charity an active user of social media and this has largely been neglected by the literature thus far.

3  DATA PROCESSING & ANALYSIS METHODS

3.1 The Alcock and Mohan data

The Alcock and Mohan data was sourced from the UK Data Archive. It contains detailed financial information for charities in England and Wales collected via a survey. The data is longitudinal and contains information from the financial years 2006/07 up to 2013/14. For this project, it was decided to use data from one financial year. The year selected was 2011/12, as the information for the more recent years was less complete and would have provided a smaller sample.

The original data, ‘CharityCharacteristics.csv’, was imported into Python as a pandas data frame. Extraneous years were then dropped from this dataframe to leave only data from 2011/12 which resulted in a sample of 12,150. Charity number was used as the index for this dataframe and is used throughout this project as it uniquely identified each charity.

The financial details in the Alcock and Mohan data are split between a large number of variables and the names used in the raw data are not particularly informative. Therefore, a ‘government funding’ variable was created by combing:

* **Government funding** ('government\_funding')   
  – combination of ig100, ig110, ig121, ig125, ig161, ig162, ig163 and ig180

Two new variables were then created to assess the proportion of each charity’s funding that came from Government Funding and General Public Funding:

* **Government Funding Proportion** (‘Prop\_government\_funding’)   
  = Government funding / total income (‘itotal’)
* **Public Funding Proportion** ('Prop\_general\_public\_funding')  
  = ig600 / total income (‘itotal’)

The data types of the variables used in these calculations were checked to ensure they were integers. Lastly, the variable for each charity’s total income (‘itotal’) was renamed as **'Income2011-2012'** to distinguish it from up-to-date information collected via a web scrape and the variable for funds generated from the general public (‘ig600’) was renamed as **'Funds\_general\_public'**.

After creation and renaming of the variables of interest, the data was filtered to create a smaller dataframe containing the 12,150 charities and 6 variables: ‘ccnum’, ‘Government\_funding', 'Funds\_general\_public', 'Prop\_government\_funding', 'Prop\_general\_public\_funding', and 'Income2011-2012'.

3.2 The Charity Commission data

It was desirable to join the Alcock and Mohan data to data scraped from the Charity Commission website for two reasons. Firstly, it contains updated income data for 2018 and secondly, it contains variables not found in the Alcock and Mohan data. Updated income allows growth to be calculated and presence in the Charity Commission data also implies survival; the Charity Commission records every active charity, so being in data from 2011-2012 (the Alcock and Mohan data) but not in data scraped in 2018 suggests that the charity is no longer operating and did not survive.

The new variables found on the Charity Commission website include ‘staff’, ‘who the charity helps’, and ‘website’. ‘Staff’ is a simple integer variable which is found in the Alcock and Mohan data but is unreliable, the version in the Charity Commission data represents fulltime equivalent employees and is collected from account returns, so is more robust. ‘Who the charity helps’ is a categorical variable which charities specify when they register from a dropdown list (it is not free-text). Example categories include ‘Children/young people’, ‘Other charities’, and ‘The general public/mankind’. We are interested in the latter, which is formatted as a binary variable in the final data. ‘Website’, where provided, is an URL to the charities own website. This was used to obtain Twitter handles as detailed in section 3.3. Several other variables were collected during the web scrape which were not used in this research but could be employed by wider projects. These include ‘Number of volunteers’, ‘Number of trustees’, ‘Expenditure’ and ‘Company number’ which could be used to link to Companies House data in the future.

The web scrape was performed using the BeautifulSoup package in Python. The Charity Commission’s website is structured so each charity’s record is identified by their unique charity number (http://beta.charitycommission.gov.uk/charity-details/?regid=***[charity number]***&subid=0). This made it easy to loop over each charity in the Alcock and Mohan data as this contained charity numbers. On each loop data was collected on all of the above variables for a single charity and these were combined into a dataframe which was saved as a JSON after all of the charities had been scraped. Error handling was built into the scraper to make it robust to charities missing data or not being registered at all, in which case all data was set to missing and the charity was recorded as not having survived.

3.3 The Twitter data

Twitter handles

In order to obtain the twitter handles for each charity it was necessary to scrape each charity's website. This was done using the Python libraries urllib and Beautiful Soup (BS). BS pulls data out of HTML files such as links to twitter.com.

The newly formed json file which contained all the previously scraped data including charity websites was imported and a pandas dataframe created. From this dataframe the websites values in the websites column were passed to the 'getTwitterHandle' function. The purpose of this function was to attempt to open the website using urlopen and, if possible, create a BS object from the html of that home page. If a BS object is created then find all instances of hyperlinks containing the word twitter and store them in a list. This list is then, in turn, iterated over to find relevant twitter links for the charity. Some websites contain links to twitter.com for reasons other than linking to their own twitter account and this should be accounted for. If the link contained certain keywords such as 'search', 'home', 'archives', 'tweet', 'share', 'intent' or 'twitter' this was likely to point to something other than the charity's twitter handle and was dropped from the list. Once a relevant link to the charity twitter handle was found the iteration was escaped and the twitter handle extracted from the link using regular expressions. The function returned the twitter handle or a placeholder '.' if no relevant twitter handle was found . This was assigned to a new column 'Twitter Handle'.

Twitter metrics

Once all the twitter handles were found the Python Tweepy library could be used to interact with the twitter api. A twitter app had to be created for the purposes of the project to obtain access tokens for permission to use the twitter api.

Similarly to the twitter handle method the dataframe was created and the Twitter Handle column was iterated over and passed to the tweepy api which had been created with a built in 'wait on rate limit' method. This was to prove invaluable as the Twitter api penalises a user who ignores the twitter api rate limit which seemed to be set at approximately 900 searches before enforcing a wait. For each twitter handle found in the previous script the tweepy api attempted to find the screen name, once the screen name was found then the metrics such as number of followers, number following and number of tweets could be easily found with inbuilt methods from the tweepy library, These were returned as pandas series with '.' place holders added for any failed users or where there were instances of having no twitter handle to check for. By the end of the iteration three more columns were created and a final Twitter info dataframe created which contained the following columns:

“ccnum”, “Twitter Handle”, “Has Twitter”, "Twitter followers", "Twitter following", “Number of tweets in total"

This was saved as a json file and would be later merged with the other files as necessary.

3.4 Combined data

The

4  ANALYSIS METHODS

4.1 Univariate methods

The

Histograms, summaries, one-way tables

4.2 Bivariate methods

The

Correlations, chi2

4.3 Multivariate methods

The

OLS

Logit

5 RESULTS & DISCUSSION

5.1 Question 1

The

5.2 Question 2: Are charities which seek to help the public more popular on Twitter?

This question investigates the link between public-facing charities and popularity on Twitter. We could assume that charities which seek to help the general public will be more heavily incentivised to engage with Twitter to interact with the public, and this would result in them being more popular on Twitter than non-public-facing charities, but to what extent to does the data substantiate this assumption?

Univariate descriptives

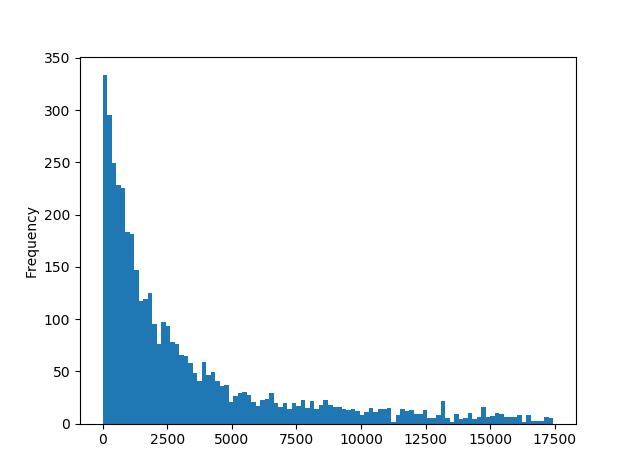
|  |  |  |
| --- | --- | --- |
|  | **Helps the general public** | **Does not help general public** |
| n | 5528 | 5272 |
| % | 51.2 | 48.8 |

The table above summarises the binary ‘Helps the general public’ variable which records if a charity seeks to help the general public or a more focused group (such as other charities, children, animals, etc.). Charities such as educational trusts are also usually excluded from this group. Roughly half of the sample is in each category which should provide good variation for modelling.

|  |  |  |  |
| --- | --- | --- | --- |
| **Does not** **help general public** | **Income 2018** | **Proportion of general public funding** | **Twitter followers** |
| **Count** | 4612 | 5747 | 1976 |
| **Mean** | 1407509 | 0.014 | 2735 |
| **Std** | 1687574 | 0.029 | 3358 |
| **Median** | 758100 | 0 | 1508 |
| **Min** | 0 | -0.046 | 0 |
| **Max** | 7400000 | 0.151 | 17324 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Helps general public** | **Income 2018** | **Proportion of general public funding** | **Twitter followers** |
| **Count** | 4846 | 5544 | 2218 |
| **Mean** | 1297389 | 0.018 | 3788 |
| **Std** | 1622773 | 0.034 | 4118 |
| **Median** | 664300 | 0 | 2134 |
| **Min** | 0 | -0.024 | 0 |
| **Max** | 7400000 | 0.151 | 17446 |

The two summary tables above are split by the binary ‘Helps the general public’ variable to show how these groups differ in terms of funding and popularity on Twitter. As shown, charities which help the general public tend to have slightly lower incomes, but gain more of their income from public funding. Most importantly, they appear to have more followers on Twitter, but this will be fully explored in the modelling.



This histogram shows the distribution of the dependent variable, twitter followers, for all charities (after it had been treated for outliers). The distribution is half-normal but is not overly afflicted with outliers and should be suitable for modelling. Followers is a good proxy for popularity because it records how many other accounts have chosen to actively subscribe to a given charity’s content on Twitter.

Multivariate modelling

|  |  |  |  |
| --- | --- | --- | --- |
| **Dependent**: Twitter followers | **Coef.** | **Std error** | **P>|t|** |
| Helps general public (binary) | 1193 | 121 | 0.000 |
| Income 2018 | 0.000 | 0.000 | 0.000 |
| Constant | 1632 | 110 | 0.000 |

R-squared 0.064 Prob = 0.000

This is an ordinary least squares model which predicts number of Twitter followers (popularity) based on whether the charity helps the general public. Income is included as a control for size which usually has a large effect on Twitter popularity (larger charities tend to be more popular on Twitter). Both the primary independent and the control are significant in this model. The primary measure is binary so the result ‘1193’ means that charities which help the general public, on average, have 1193 more twitter followers than those who do not – controlling for size. This suggests public facing charities are more popular on Twitter and positively affirms research question 2. However, the R-squared for this model is small which suggests there are many other factors (or simply random variation) which affect charity popularity on Twitter.

5.3 Question 3: Does number of staff, rather than size, determine a charity’s active use of Twitter?

Summary statistics

In order to determine whether the number of staff employed by a charity, rather than the size of the charity, plays a role in a charity’s successful use of Twitter, three variables were selected from the final data set: ‘Staff’, ‘Twitter following’, and ‘Income2018’. ‘Twitter following’ is the number of accounts the charity follows and is used as a measure of active use of Twitter, as some charities may choose not to Tweet very often but may still follow other accounts which shows a level of use.

**Table 1: Summary statistics for Staff, Twitter following and Income2018**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Staff** | **Twitter following** | **Income2018** |
| **Data type** | float64 | float64 | float64 |
| **Count** | 10695.000000 | 4454.000000 | 9.458000e+03 |
| **Mean** | 14.729874 | 1000.057252 | 1.351087e+06 |
| **Standard deviation** | 23.886958 | 885.686228 | 1.655516e+06 |
| **Minimum** | 0.000000 | 0.000000 | 0.000000e+00 |
| **25%** | 0.000000 | 289.000000 | 1.382500e+05 |
| **50%** | 0.000000 | 758.500000 | 7.067000e+05 |
| **75%** | 22.000000 | 1505.500000 | 1.900000e+06 |
| **Maximum** | 102.000000 | 3874.000000 | 7.400000e+06 |

Bivariate Correlation

The correlation between the number of Staff and the number ‘Following’ on Twitter was calculated using df['Staff'].corr(df['Twitter following'] and a weak positive correlation of 0.07 was measured.

Multivariate modelling

A linear regression was carried out (y = m1x1 + m2x2) using ‘Staff’ and ‘Income2018’ as the independent variables: x1 and x2, and setting ‘Twitter following’ as the dependent variable: y. The statmodels.api module was used and the results are shown in Table 2.

**Table 2: Results of the Linear Regression Analysis**

The Adjusted R-squared value takes into account the number of independent variables used in the model and in this case it is 0.016. The constant coefficient is the intercept with the y axis and is 867.4814. The standard error indicates the accuracy of the constant, ‘Staff’ and ‘Income2018 coefficients with a low value indicating a higher accuracy, and the confidence interval shows the range that the coefficients will probably fall within.

The p-value of 0.189 for ‘Staff’ indicates that changes in this variable are not associated with changes in ‘Twitter following’ – the y variable. However, the p-value of 0 for ‘Income2018’ is statistically significant, so the null hypothesis that size of a charity does not affect the number of people / organisations that the charity follows on Twitter can be rejected. Note: Size of the charity is measured by the Income in this case.

6 CONCLUSIONS

6.1 Summary

The

6.2 Future work

The

REFERENCES