Predicting popularity of categories in Deal Related Tweets

Project Report : Social Media Mining

Tanmoy Choudhury
Indiana University
Bloomington, IN
tdchoudh@iu.edu

Madhavi Pilli
Indiana University
Bloomington, IN
mpilli@iu.edu

Vipin Singh
Indiana University
Bloomington, IN
vipsingh@iu.edu

ABSTRACT

Popularity of deal related tweets is an indication of the reach of twitter users to their targeted consumers in marketing their deals. This research aims at predicting popularity of the categories of deal related tweets. Using hash tags to extract tweets related to deals we categorized the tweets in to 5 broad categories and tried to predict popular deal categories based on its features. We tried to prove correlation of likes and re-tweets on the popularity of the category. Based on the data and research question using Linear Regression and SVM machine learning algorithms proved the best choice for this research problem. Results do not indicate strong correlation between like, re-tweet and the category of the deal, however we have interesting observations that could aid in further research in this space.

1. INTRODUCTION

Twitter has been used in a lot of research areas in business, academia and health since its launch in 2006 wiki(Wikipedia, 2017). It is one of the popular micro-blogging sites that provides a medium for users to communicate in less than 280 characters after the recent upgrade except for few languages. Businesses use this social media network to advertise their deals to reach their consumers. Users use twitter and twitter based applications to save money using coupons and deals. There are certain twitter accounts that specifically work on collating and sharing deals with their followers. In this research, we plan on analyzing deal related tweets to find features that could be impacting the popularity of the deal category. This is very interesting topic based on the fact that deal is a very catchy word and it involves saving money. Average consumers are always on the outlook for deals on several items they plan to buy each year and the crowd we see during thanksgiving is a good enough evidence to prove this.

2. RELATED WORKS

There are numerous research studies related to twitter classification in certain categories like social marketing, fitness, disaster, trending etc.

Yu, Chen, and Kwok (2011) work on a similar study that tries to predict popularity of Social Marketing messages on Facebook based on its features like content and media type. Support Vector Machine (SVM) and Naive Bayes algorithms were used in this research and we see that SVM has proved to be efficient in text mining use cases. We also used SVM for this research, however our feature selection needs to be improved further.

Theodotou and Stassopoulou (2015) proposed a system to automatically categorize tweets into different classification. Furthermore, the system considers additional features like linked URL, user profiles and articles from Wikipedia. The model achieved higher accuracy when these feature set was used. The implementation is done using C# twitter api with specific hash tags to categorize the tweets. Once the data is collected, Naive Bayes classifier proved more accurate in this case for feature group having user profile and URL terms.

Vickey, Ginis, and Dabrowski (2013) explored fitness related tweet to identify usages of fitness related apps. The data is collected through an open source program called TwapperKeeper, which is now part of HootSuite. Additionally, mobile fitness app related hash tags were used. A computerized text classification procedure was used to identify activity and conversation tweets.This provides an insight into location based workout and exercise frequency. Similarly, future research in this area could be based on building a classification model that can provide most popular deal categories along with geo-location info.

Dayani, Chhabra, Kadian, and Kaushal (2015) studied to identify reliability of a tweet. There could be rumour tweets which may be unreliable and when they get re-tweeted, it becomes widespread. This can be included as a future work for our study. we could benefit adding a method for rumor detection to our work to ensure originality of tweets as a pre-processing step.

Becker et al.Becker, Naaman, and Gravano (2011) explored an end to end approach to identify trending events. Each event and its associated messages are grouped together into similar tweets using a clustering algorithm. The study also collated similar features for each cluster and use them to train classifier to differentiate between event and non-event cluster. The article did not explain further regarding the specific method used.

Sriram, Fuhry, Demir, Ferhatosmanoglu, and Demirbas (2010) proposed a small set of domain specific feature set extraction from author's profile and text rather than using standard classification method such as ”Bag-Of-Words”. The proposed approach classifies incoming tweet into categories like News, Opinions and Deals. This article does not provide popularity among various deals category and our research aims at achieving this.

2.1 Our Contribution

As we see several studies around social media marketing analysis and predictions targeting Facebook and twitter, one area that could benefit from our research is the popularity prediction of Deals related tweets. Our study targets finding correlation of features like, likes and re-tweets on popularity of deal category. We limited our research to 5 categories of deals to start our research, however this can easily be extended to cover other categories as well.

3. DATA ACQUISITION

Acquiring tweets, especially related to deals is our area of interest for this research. We targeted acquiring tweets for the year 2016 to confine our research dataset for the purpose of this project. This is because 2016 is the previous full year at the time of this project and the one year confinement is keeping project time lines and software we are using in mind. Official twitter API has a constraint that restricts getting tweets older than a week. Due to this limitation we used got3 based on lxml and pyquery. Got3 is implemented based on twitter search through browser. Using this we extracted tweets of our interest, in this case deal related tweets by limiting to 5000 tweets that matches our search criteria. We initially tried getting all deal related tweets for each day and ended up with almost 30K tweets each day. Due to time and process limitations we had for this project we set 5000 tweets limit per day and ran our process to extract tweets into a csv file one for each day. Challenge we ran into while acquiring data is defining our search criteria to get the tweets of our interest. Our search criteria was based on the hashtags - deals, discounts, sale, promotion, save, dealoftheday. Our initial criteria included several other hashtags. Using these complete set of hashtags we ran into issues with data volumes and hence decided to limit our hashtags to these 6 along with a maxlimit of 5000 tweets per day. With all these limits we still had a very good spread of data across the 5 categories to perform our prediction.

4. DATA PRE-PROCESSING

Our pre-processing step involved traversing through more than 360 files we had collected as part of our data acquisition to extract additional features needed for machine learning as well as cleansing the data. This can be divided into two subsections. Firstly, we combine all individual files to a single csv file and cleanse the tweet text. Secondly, we do a word matching to identify individual categories.

4.1 Data Cleansing

As part of pre-processing, the tweet text we tried to extract an additional feature called pic indicator to indicate if a picture is included in the tweet. This was done in anticipation to see if this feature had any correlation to the popularity. Data cleansing involved converting the text to lower case, removing numbers, white spaces, tickers, stop words and repetitive characters. Date field of the tweet is processed to derive month as well as the quarter the message was tweeted.

4.2 Tweet Categorization

Deal related tweet data is analyzed to come up with features that could be extracted for our research. Initial exploratory analysis helped in coming up with search criteria. We chose five broad categories. They are “Electronics”,“Clothing”,“Health and Beauty”,“Book”, “Home”. Tweets which does not fall within these categories are bucketed into “Other”. For each category we manually identified certain set of words. We searched each and every tweet for a string match and assign appropriate category to individual tweet. Tweet text which does not match with any of the words for category are defaulted to “Other”. The words assigned to each category are represented in TABLE 1 below.

Electronics	cellular,phone,iphone,ipad,samsung, desktop,laptop,dell,tv,television,xbox, camera,printer,wireless,games,dslr,intel, notebook,playstation,usb,smartphone, macbook,tablet,android
Clothing	shirt,denim,dress,sweater,coat,apparel, jacket,clothing,sweatshirt,polos,fashion, leggings,socks,jogger, jeans, pant, bodysuit,onesie,pajama,wool,suit, shirts, pullover,t-shirt,gloves,hat,skirt, pants,size
Health and Beauty	fitness,silver,gold,diamond,ring, nail,perfume,bracelet,weight,loss, diet,brooch,earrings,health, bodybuilding,body,fashion,health, solitaire,yogamat,skincare,toilette, massage,sapphire,beauty,style,hair, boutique,yoga,candle,crystal,jewelry, pendant
Book	book,fiction,non-fiction,fict,nonfict, bestseller,bestselling,books,ebook, kindle, comics, story, novel,author, bookmark, pen, stationary
Home	grilling,microwave,vaccum,cleaner, vanity,stool,furniture,dining,kitchen, table,chair,sofa,cookware,calphalon, mugs,tourister, skybags, luggage, garden,refrigerator,washing, dishwasher,light,house,fridge,decor, homedecor,bbq,pot,storage, portrait,paint

Table 1: Words used in Tweet Categorization

5. DATA ANALYSIS

Exploring the data we had processed and categorized led to some interesting observations. We wanted to see trending across different months in the year, see the distribution of tweets across different categories we had chosen as well as the visualization of dependent variables(likes, retweet count).

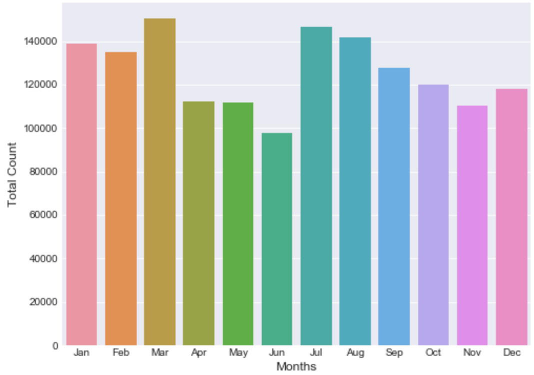


Figure 1: Tweet by Month

Figure 1 illustrates deal related tweet data collected over 2016. The count of tweet varies across months but averaged near 100 thousands. The tweet volume is more at the first quarter of 2016 and slightly drops in the second quarter. It picks up again on 3rd quarter and becomes steady till end of year.

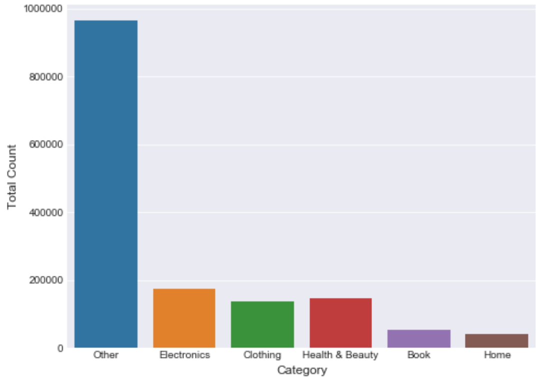


Figure 2: Count of Tweet by Category

Figure 2 indicates count of tweet by category for the entire year 2016. As part of this research we have categorize deal related tweet into five different category and everything else into others. We do see that majority of the tweets were in "Others" i.e non categorized forms. For our research question we will remove "Other" category.

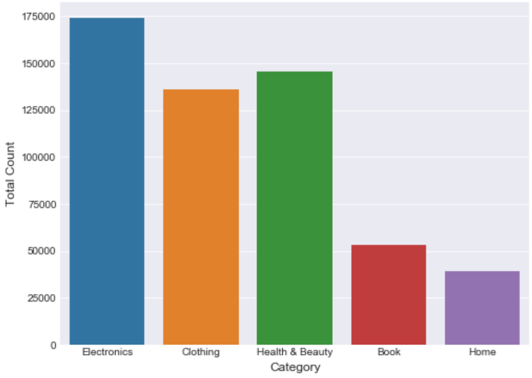


Figure 3: Count of Tweet by Category Excluding Other

Figure 3 represents tweet by category excluding others. Among five different categories we see majority of them belong to "Electronics". Around 175 thousands tweets were categorized as "Electronics". Next is "Health" and "Beauty" in around 150 thousands. "Clothing" is slightly over 125 thousands and the remaining categories "Book" and "Home" are below 50 thousands.

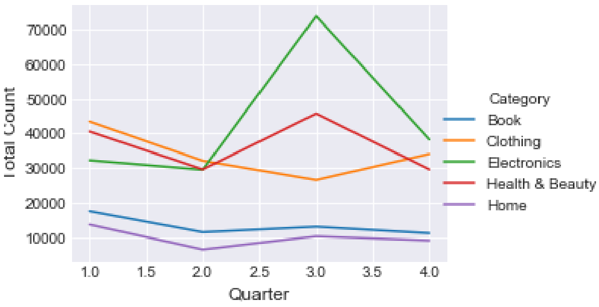


Figure 4: Tweet Category By Quarter

Figure 4 shows tweet category by quarter. We see a steady increase for "Electronics" category from second quarter till middle of third quarter and sharp decline thereafter till quarter four. Even though there is a decline on "Electronics" tweet the average count is more than the first quarter. "Health and Beauty" shows similar pattern. This kind of proves that "Electronics" and "Health and Beauty" are popular category as Thanksgiving and year end approaches. However, deal tweets related to "Clothing" stars higher at the start of the year and gradually decline till end of second quarter and steady increase thereafter for the holiday season. Out of these five categories only two category "Book" and "Home" has relatively low tweet volume and not much variance over the year.

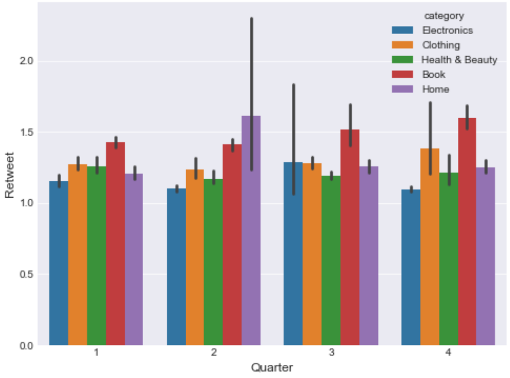


Figure 5: Retweet By Quarter

Figure 5 shows retweet by category over the quarter. Apart from second quarter, “Book” and “Clothing” are mostly re tweeted.Even though the count of tweet for “Book” category are far less, they are most re tweeted.

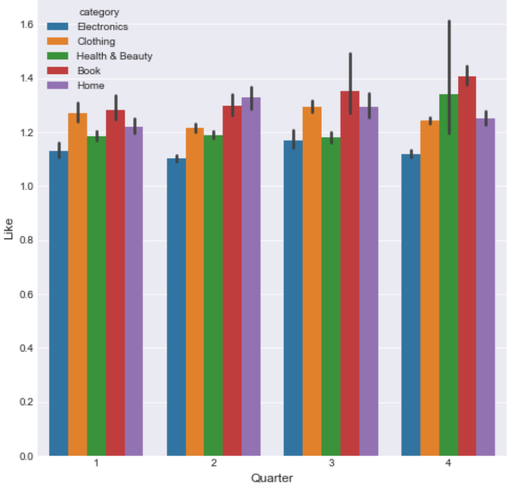


Figure 6: Like By Quarter

Figure 6 illustrate tweet likes by category over the quarter. Similar to Figure 5, “Book” and “Clothing” are mostly liked apart from second quarter.

6. RESEARCH METHODS

Our goal is to identify highly influential features for deal tweets. Features, as per the collected data set, are date, month, day, pic_ind and various categories. We decided to tackle the problem by applying two approaches. First method we used is Multiple Linear regression, two equations were created where like and retweet are the dependent variables and feature set is the independent variable. We proceeded to reject null hypothesis by proving that a multiple linear relationship exists between dependent and independent variable. Multiple Linear regression is opted because this method is very effective in solving this type of problems. The specification test makes it sluttish to observe the coefficients of various features along with statistical significance of each independent variable.Statistical strength of any variable in specification test is represented by pvalue. If pvalue is less than 0.005, then the variables are considered as statistically strong. A substantial R-squared and adjusted R-squared facilitates to reject null hypothesis. The summary table generated for Multiple Linear Regression, is ascertained for positive and negative impingement of various independent variable on dependent variables. We noted the relations ship on log of independent variables as well. Since log of 0 is not defined, we have increased the count of like and re-tweet by 1 i.e if like is 0 it is made as 1, 1 made as 2 and so on.

Additionally, we decided to discover influencing features by treating this problem as a classification problem. A new column ”popular tweet” is appended to dataset to define popular and non popular tweets. The value 1 represented popular and value 0 equated to non popular tweets. This column is populated by verifying the quantiles of like and retweet dependent variables. A tweets is considered as popular if like dependent variable is greater than or equal to 97th percentile. Hence all like having a value of 3 (corresponds to 97th percentile, considered 97 for like and 98 for retweet) or more are considered as popular tweets, others as non popular. Support Vector Machine and Logistic Regression were used to solve this binary classification problem. The algorithms were imported from sklearn python library. We knew that SVC takes long time to run but we valued it to use because it uses kernel trick and it has a regularization parameter.Logistic regression has a low variance and it has very less chances of over-fitting. These characteristics makes these two algorithms a good choice to get better accuracy. A random baseline accuracy is very crucial statistic to recognize, before using any algorithm to solve classification problem. The accuracy of algorithm and random baseline is compared to evaluate the difference. A poor output of algorithm shows a very less difference from random baseline.The Logistic Regression and SVM were used to predict popular and non popular tweets. Using popular tweets we identified various popular categories.Finally we checked the weights of the words of tweets with popular categories and listed highest weighted words in popular tweets.

7. RESULTS

Multiple Linear Regressions are used to determine the relationship between independent and dependent variables. We designed the following multiple linear equations for like(1), retweet(2) and logarithmic versions log(like) (3) and log(retweet) (4) based on our assumptions.:

$$\text{like} = \beta_0 + \beta_1\ \text{elec} + \beta_2\ \text{cloth} + \beta_3\ \text{healthbeauty} + \beta_4\ \text{book} + \beta_5 \text{home} + \beta_6\ \text{wday} + \beta_7\ \text{wend} + \beta_8\ \text{jan} + \beta_9\ \text{feb} + \beta_{10}\ \text{mar} + \beta_{11}\ \text{apr} + \beta_{12}\ \text{may} + \beta_{13}\ \text{jun} + \beta_{14}\ \text{jul} + \beta_{15}\ \text{aug} + \beta_{16}\ \text{sep} + \beta_{17}\ \text{oct} + \beta_{18}\ \text{nov} + \beta_{19}\ \text{dec} + \beta_{20}\ \text{char} + \mu$$

$$\text{retweet} = \beta_0 + \beta_1\ \text{elec} + \beta_2\ \text{cloth} + \beta_3\ \text{healthbeauty} + \beta_4\ \text{book} + \beta_5 \text{home} + \beta_6\ \text{wday} + \beta_7\ \text{wend} + \beta_8\ \text{jan} + \beta_9\ \text{feb} + \beta_{10}\ \text{mar} + \beta_{11}\ \text{apr} + \beta_{12}\ \text{may} + \beta_{13}\ \text{jun}$$

$$+ \beta_{14} \text{ jul} + \beta_{15} \text{ aug} + \beta_{16} \text{ sep} + \beta_{17} \text{ oct} + \beta_{18} \text{ nov} + \beta_{19} \text{ dec} + \beta_{20} \text{ char} + \mu$$

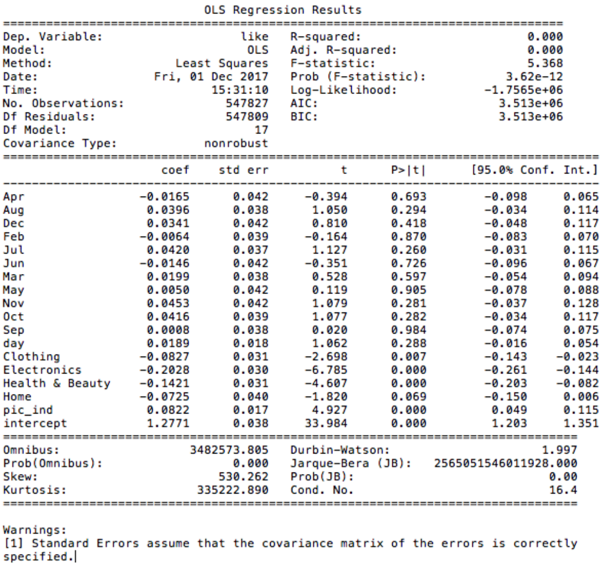
Logarithmic version:

$$\log(\text{like}) = \beta_0 + \beta_1 \text{ elec} + \beta_2 \text{ cloth} + \beta_3 \text{ healthbeauty} + \beta_4 \text{ book} + \beta_5 \text{home} + \beta_6 \text{ wday} + \beta_7 \text{ wend} + \beta_8 \text{ jan} + \beta_9 \text{ feb} + \beta_{10} \text{ mar} + \beta_{11} \text{ apr} + \beta_{12} \text{ may} + \beta_{13} \text{ jun} + \beta_{14} \text{ jul} + \beta_{15} \text{ aug} + \beta_{16} \text{ sep} + \beta_{17} \text{ oct} + \beta_{18} \text{ nov} + \beta_{19} \text{ dec} + \beta_{20} \text{ char} + \mu$$

$$\log(\text{retweet}) = \beta_0 + \beta_1 \text{ elec} + \beta_2 \text{ cloth} + \beta_3 \text{ healthbeauty} + \beta_4 \text{ book} + \beta_5 \text{home} + \beta_6 \text{ wday} + \beta_7 \text{ wend} + \beta_8 \text{ jan} + \beta_9 \text{ feb} + \beta_{10} \text{ mar} + \beta_{11} \text{ apr} + \beta_{12} \text{ may} + \beta_{13} \text{ jun} + \beta_{14} \text{ jul} + \beta_{15} \text{ aug} + \beta_{16} \text{ sep} + \beta_{17} \text{ oct} + \beta_{18} \text{ nov} + \beta_{19} \text{ dec} + \beta_{20} \text{ char} + \mu$$

A null hypothesis states that there is no linear relations ship exists between independent and dependent variables.We commenced with the process of rejecting null hypothesis by observing the values of R-Squared and adjusted R squared. We kickedoff with 547827 tweets and performed a specification test on both equations. Initial results revealed most of the coefficients as same and R-Squared as 0. We realized that there is some issue with feature set. Further investigations in feature set indicated a problem of multicollinearity in our independent features. Multi Collinearity occurs if one or more feature set are highly collinear to each other. In order to get the specification test veracious, we reduced the collinearity by removing "January" and "Book" features and also introduced the intercept in the model."January" feature removal has reduced the collinearity in the group of months and "Book" in the group of categories. January was collinear with other eleven months and similarly Book was collinear with other categories. Hence it was easy for the system to derive the "January" feature by not considering other eleven months and similarly the "Book" feature can be deduced by not considering other categories. We used stats model in python to acquire p-values to understand the statistical implications on like and retweet. The eliminated features "Jan" and "Book" as a part of specification test will contribute to intercept or default. We also did the specification test on the log of dependent variable because of the skewed nature of the data set and illustrated the results.

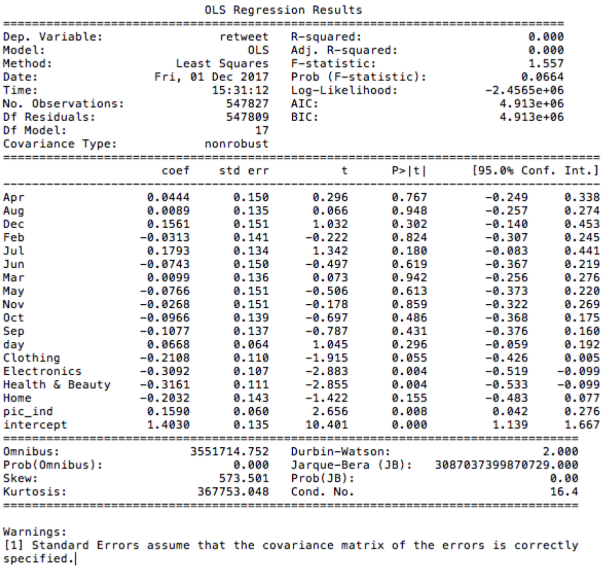
The results of first two equations were not able to reject null hypothesis. We observed that RSquared and Adjusted RSquared values were 0.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   like   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     5.368
Date:                Fri, 01 Dec 2017   Prob (F-statistic):           3.62e-12
Time:                        15:31:10   Log-Likelihood:            -1.7565e+06
No. Observations:              547827   AIC:                         3.513e+06
Df Residuals:                  547809   BIC:                         3.513e+06
Df Model:                          17
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------
Apr               -0.0165      0.042     -0.394      0.693        -0.098     0.065
Aug                0.0396      0.038      1.050      0.294        -0.034     0.114
Dec                0.0341      0.042      0.810      0.418        -0.048     0.117
Feb               -0.0064      0.039     -0.164      0.870        -0.083     0.070
Jul                0.0420      0.037      1.127      0.260        -0.031     0.115
Jun               -0.0146      0.042     -0.351      0.726        -0.096     0.067
Mar                0.0199      0.038      0.528      0.597        -0.054     0.094
May                0.0050      0.042      0.119      0.905        -0.078     0.088
Nov                0.0453      0.042      1.079      0.281        -0.037     0.128
Oct                0.0416      0.039      1.077      0.282        -0.034     0.117
Sep                0.0008      0.038      0.020      0.984        -0.074     0.075
day                0.0189      0.018      1.062      0.288        -0.016     0.054
Clothing          -0.0827      0.031     -2.698      0.007        -0.143    -0.023
Electronics       -0.2028      0.030     -6.785      0.000        -0.261    -0.144
Health & Beauty   -0.1421      0.031     -4.607      0.000        -0.203    -0.082
Home              -0.0725      0.040     -1.820      0.069        -0.150     0.006
pic_ind            0.0822      0.017      4.927      0.000         0.049     0.115
intercept          1.2771      0.038     33.984      0.000         1.203     1.351
==============================================================================
Omnibus:                  3482573.805   Durbin-Watson:                   1.997
Prob(Omnibus):                  0.000   Jarque-Bera (JB):   2565051546011928.000
Skew:                         530.262   Prob(JB):                         0.00
Kurtosis:                  335222.890   Cond. No.                         16.4
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Figure 7: Multiple Linear Regression of Like

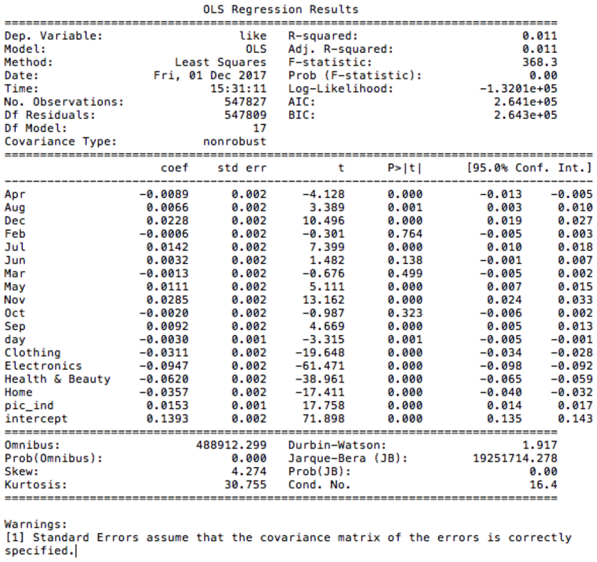
Figure 7 shows regression results for like. The R-Squared and adjusted Rsquared are 0. If we observe closely only the pic_ind has a positive impact on the like and it is statistically strong having a pvalue of 0, the categories clothing, electronics, health beauty and home are also statistically significant, but have a negative impact on like. We have removed the Jan and Book features and both features have contributed to default which shows a 0.4 (because a 1 intercept was added) coefficient of positive impact.The default also shows statistically significant.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                retweet   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     1.557
Date:                Fri, 01 Dec 2017   Prob (F-statistic):             0.0664
Time:                        15:31:12   Log-Likelihood:            -2.4565e+06
No. Observations:              547827   AIC:                         4.913e+06
Df Residuals:                  547809   BIC:                         4.913e+06
Df Model:                          17
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------
Apr                0.0444      0.150      0.296      0.767        -0.249     0.338
Aug                0.0089      0.135      0.066      0.948        -0.257     0.274
Dec                0.1561      0.151      1.032      0.302        -0.140     0.453
Feb               -0.0313      0.141     -0.222      0.824        -0.307     0.245
Jul                0.1793      0.134      1.342      0.180        -0.083     0.441
Jun               -0.0743      0.150     -0.497      0.619        -0.367     0.219
Mar                0.0099      0.136      0.073      0.942        -0.256     0.276
May               -0.0766      0.151     -0.506      0.613        -0.373     0.220
Nov               -0.0268      0.151     -0.178      0.859        -0.322     0.269
Oct               -0.0966      0.139     -0.697      0.486        -0.368     0.175
Sep               -0.1077      0.137     -0.787      0.431        -0.376     0.160
day                0.0668      0.064      1.045      0.296        -0.059     0.192
Clothing          -0.2108      0.110     -1.915      0.055        -0.426     0.005
Electronics       -0.3092      0.107     -2.883      0.004        -0.519    -0.099
Health & Beauty   -0.3161      0.111     -2.855      0.004        -0.533    -0.099
Home              -0.2032      0.143     -1.422      0.155        -0.483     0.077
pic_ind            0.1590      0.060      2.656      0.008         0.042     0.276
intercept          1.4030      0.135     10.401      0.000         1.139     1.667
==============================================================================
Omnibus:                  3551714.752   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):   3087037399870729.000
Skew:                         573.501   Prob(JB):                         0.00
Kurtosis:                  367753.048   Cond. No.                         16.4
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Figure 8: Multiple Linear Regression of Retweet

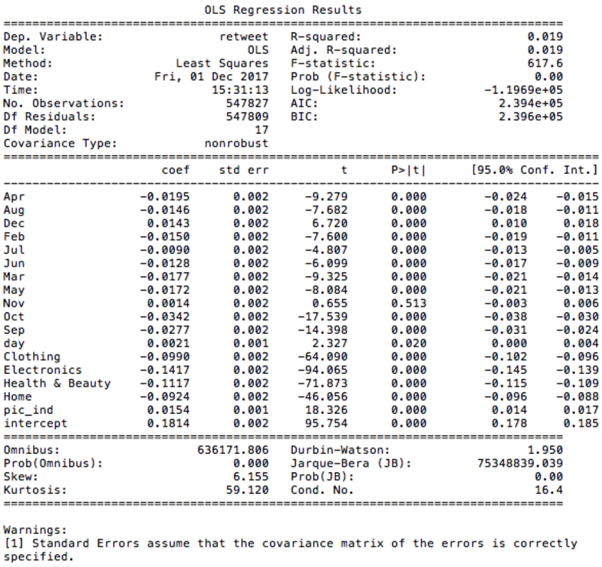
Figure 8 shows regression results for retweet. The R-Squared and adjusted Rsquared are 0. If we observe closely only the pic_ind has a positive impact on the like and it is less statistically strong compared with like, the categories electronics and health beauty are also statistically pregnant, but have a negative impact on retweet. We have removed the Jan and Book features and they have contributed to default which shows a 0.4 coefficient meaning a 0.6 negative affect (intercept was added as 1). The default also shows high statistically importance.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   like   R-squared:                       0.011
Model:                            OLS   Adj. R-squared:                  0.011
Method:                 Least Squares   F-statistic:                     368.3
Date:                Fri, 01 Dec 2017   Prob (F-statistic):               0.00
Time:                        15:31:11   Log-Likelihood:            -1.3201e+05
No. Observations:              547827   AIC:                         2.641e+05
Df Residuals:                  547809   BIC:                         2.643e+05
Df Model:                          17
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [95.0% Conf. Int.]
-----------------------------------------------------------------------------------
Apr                -0.0089      0.002     -4.128      0.000        -0.013    -0.005
Aug                 0.0066      0.002      3.389      0.001         0.003     0.010
Dec                 0.0228      0.002     10.496      0.000         0.019     0.027
Feb                -0.0006      0.002     -0.301      0.764        -0.005     0.003
Jul                 0.0142      0.002      7.399      0.000         0.010     0.018
Jun                 0.0032      0.002      1.482      0.138        -0.001     0.007
Mar                -0.0013      0.002     -0.676      0.499        -0.005     0.002
May                 0.0111      0.002      5.111      0.000         0.007     0.015
Nov                 0.0285      0.002     13.162      0.000         0.024     0.033
Oct                -0.0020      0.002     -0.987      0.323        -0.006     0.002
Sep                 0.0092      0.002      4.669      0.000         0.005     0.013
day                -0.0030      0.001     -3.315      0.001        -0.005    -0.001
Clothing           -0.0311      0.002    -19.648      0.000        -0.034    -0.028
Electronics        -0.0947      0.002    -61.471      0.000        -0.098    -0.092
Health & Beauty    -0.0620      0.002    -38.961      0.000        -0.065    -0.059
Home               -0.0357      0.002    -17.411      0.000        -0.040    -0.032
pic_ind             0.0153      0.001     17.758      0.000         0.014     0.017
intercept           0.1393      0.002     71.898      0.000         0.135     0.143
==============================================================================
Omnibus:                   488912.299   Durbin-Watson:                   1.917
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         19251714.278
Skew:                           4.274   Prob(JB):                         0.00
Kurtosis:                      30.755   Cond. No.                         16.4
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Figure 9: Multiple Linear Regression of log(like)

Figure 9 shows regression results for log of like and shows better results compare to like linear regression. The R-Squared and adjusted Rsquared are greater than 0, but the model still performs poorly. This value of 0.011 is not enough to reject null hypothesis. If we observe closely this regression shows more statistically strong variables. The pvalue is less than 0.005 for apr, dec, jul, may and nov in terms of months and shows a positive impact on log(like). All categories seems to impact log(like) in a positive way.



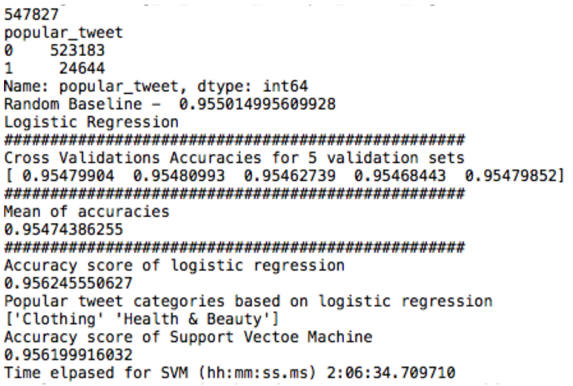
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                retweet   R-squared:                       0.019
Model:                            OLS   Adj. R-squared:                  0.019
Method:                 Least Squares   F-statistic:                     617.6
Date:                Fri, 01 Dec 2017   Prob (F-statistic):               0.00
Time:                        15:31:13   Log-Likelihood:            -1.1969e+05
No. Observations:              547827   AIC:                         2.394e+05
Df Residuals:                  547809   BIC:                         2.396e+05
Df Model:                          17
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [95.0% Conf. Int.]
-----------------------------------------------------------------------------------
Apr                -0.0195      0.002     -9.279      0.000        -0.024    -0.015
Aug                -0.0146      0.002     -7.682      0.000        -0.018    -0.011
Dec                 0.0143      0.002      6.720      0.000         0.010     0.018
Feb                -0.0150      0.002     -7.600      0.000        -0.019    -0.011
Jul                -0.0090      0.002     -4.807      0.000        -0.013    -0.005
Jun                -0.0128      0.002     -6.099      0.000        -0.017    -0.009
Mar                -0.0177      0.002     -9.325      0.000        -0.021    -0.014
May                -0.0172      0.002     -8.084      0.000        -0.021    -0.013
Nov                 0.0014      0.002      0.655      0.513        -0.003     0.006
Oct                -0.0342      0.002    -17.539      0.000        -0.038    -0.030
Sep                -0.0277      0.002    -14.398      0.000        -0.031    -0.024
day                 0.0021      0.001      2.327      0.020         0.000     0.004
Clothing           -0.0990      0.002    -64.090      0.000        -0.102    -0.096
Electronics        -0.1417      0.002    -94.065      0.000        -0.145    -0.139
Health & Beauty    -0.1117      0.002    -71.873      0.000        -0.115    -0.109
Home               -0.0924      0.002    -46.056      0.000        -0.096    -0.088
pic_ind             0.0154      0.001     18.326      0.000         0.014     0.017
intercept           0.1814      0.002     95.754      0.000         0.178     0.185
==============================================================================
Omnibus:                   636171.806   Durbin-Watson:                   1.950
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         75348839.039
Skew:                           6.155   Prob(JB):                         0.00
Kurtosis:                      59.120   Cond. No.                         16.4
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Figure 10: Multiple Linear Regression of log(retweet)

Figure 10 shows regression results for log of retweet and shows better results compare to retweet linear regression. The R-Squared and adjusted Rsquared are greater than 0, but the model still performs poorly. This value of 0.019 is not enough to reject the null hypothesis. If we observe closely this regression shows more statistically satisfying variables. The pvalue is greater than 0.005 for nov and day. It means the month "November" and day variable which can be "weekday" or "weekend" are not that statistically substantial. Other variables shows negative impact on log of retweet except the month of "December" and pic_ind.

Observing all four regression tables above, none of the regression was able to reject null hypothesis and hence we could not prove that there is linear relationship exists as per our assumption.

Since linear regression did not demonstrated promising results, we took another approach to analyze this data, and tried to solve it by using classification algorithms. The 547827 tweets were split in to popular and non popular tweets. The tweets where like or retweet is in more than 97th and 98th percentile respectively were considered as popular tweets and rest were considered as non popular tweets. We have used python sklearn to obtain results. The split gave us 24644 as popular tweets and 523183 as non popular tweets. We used TfidfVectorizer to transform tweet text in to numbers to feed in to classification algorithms. The column popular_tweet is considered as the classification output variable.Random baseline is calculated on the split dataset by using the formula: (523183/(523183 + 24644)) = 0.955014995609928.



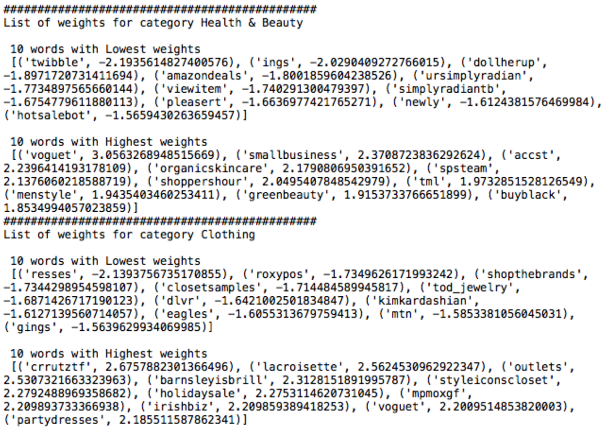
```
547827
popular_tweet
0    523183
1     24644
Name: popular_tweet, dtype: int64
Random Baseline -  0.955014995609928
Logistic Regression
####################################################
Cross Validations Accuracies for 5 validation sets
[ 0.95479904  0.95480993  0.95462739  0.95468443  0.95479852]
####################################################
Mean of accuracies
0.95474386255
####################################################
Accuracy score of logistic regression
0.956245550627
Popular tweet categories based on logistic regression
['Clothing' 'Health & Beauty']
Accuracy score of Support Vectoe Machine
0.956199916032
Time elpased for SVM (hh:mm:ss.ms) 2:06:34.709710
```

Figure 11: Classification Algorithm output along with random baseline

Figure 11 shows the split, output of random baseline and classification algorithms output of Logistic Regression and Support Vector Machine.A cross validation score of logistic regression along with the mean is also shown. Observing it closely the mean of cross validation score is slightly less than the random baseline score. It means the data set is not training the algorithm effectively. We expect an accuracy higher than the random baseline. We tried to classify the data by using support vector machine and the output is displayed, the accuracy of SVM is slightly better than the baseline. The classification algorithms output is still poor, but better than logistic regression.

We then predicted the categories of the popular tweets and found that Clothing and Health Beauty are the popular categories among the total tweets.

Finally we found out the words with highest and lowest weights in popular categories.



```
##############################################
List of weights for category Health & Beauty

 10 words with Lowest weights
 [('twibble', -2.1935614827400576), ('ings', -2.0290409272766015), ('dollherup',
-1.8971720731411694), ('amazondeals', -1.8001859604238526), ('ursimplyradian',
-1.7734897565660144), ('viewitem', -1.740291300479397), ('simplyradiantb',
-1.6754779611880113), ('pleasert', -1.6636977421765271), ('newly', -1.6124381576469984),
('hotsalebot', -1.5659430263659457)]

 10 words with Highest weights
 [('voguet', 3.0563268948515669), ('smallbusiness', 2.3708723836292624), ('accst',
2.2396414193178109), ('organicskincare', 2.1790806950391652), ('spsteam',
2.1376060218588719), ('shoppershour', 2.0495407848542979), ('tml', 1.9732851528126549),
('menstyle', 1.9435403460253411), ('greenbeauty', 1.9153733766651899), ('buyblack',
1.8534994057023859)]
##############################################
List of weights for category Clothing

 10 words with Lowest weights
 [('resses', -2.1393756735170855), ('roxypos', -1.7349626171993242), ('shopthebrands',
-1.7344298954598107), ('closetsamples', -1.714484589945817), ('tod_jewelry',
-1.6871426717190123), ('dlvr', -1.6421002501834847), ('kimkardashian',
-1.6127139560714057), ('eagles', -1.6055313679759413), ('mtn', -1.5853381056045031),
('gings', -1.5639629934069985)]

 10 words with Highest weights
 [('crrutztf', 2.6757882301366496), ('lacroisette', 2.5624530962922347), ('outlets',
2.5307321663323963), ('barnsleyisbrill', 2.3128151891995787), ('styleiconscloset',
2.2792488969358682), ('holidaysale', 2.2753114620731045), ('mpmoxgf',
2.209893733366938), ('irishbiz', 2.209859389418253), ('voguet', 2.2009514853820003),
('partydresses', 2.185511587862341)]
```

Figure 12: Highest and Lowest weights of popular categories

Figure 12 shows weights of category clothing and Health and Beauty.The words like ”lacroisette” and ”outlets” have the highest weights in clothing and the words like ”voguet”, ”smallbusiness” and ”organicskincare” have highest weights.

8. CONCLUSION

As per the results section we have tried regression and classification to actually solve the problem of predicting popular categories in deal related tweets. The regession was not able reject the null hypothesis, hence we did further data exploration to get more insights in the data.

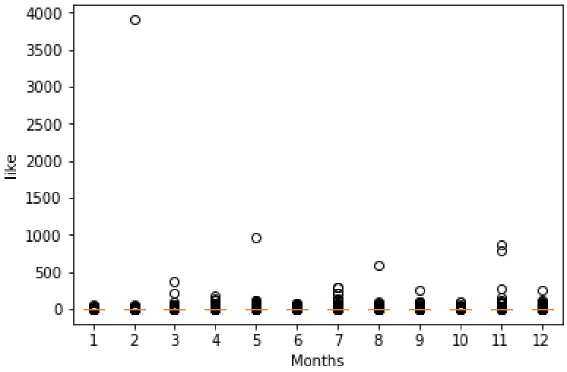


Figure 13: Distribution of like on months

Figure 13 shows distribution of like on months. It seems that the distribution is close to normal and any specific month does not show the influence on like. Hence the like regression showed a 0 RSquared, because actually there is no linear relationship exists between months and likes.

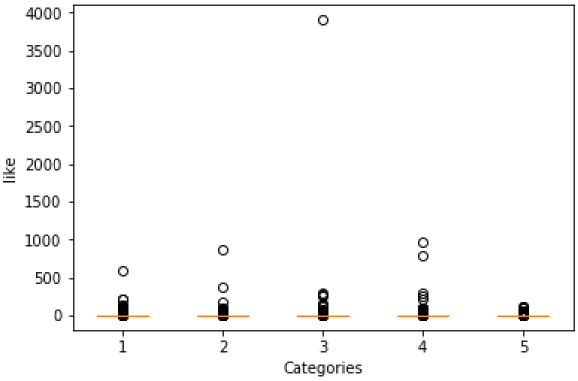


Figure 14: Distribution of like on Categories

Figure 14 shows distribution of like on categories. It seems that the distribution is close to normal and any specific category does not show the influence on like. Hence the like regression showed a 0 RSquared, because actually there is no linear relationship exists between categories and likes.

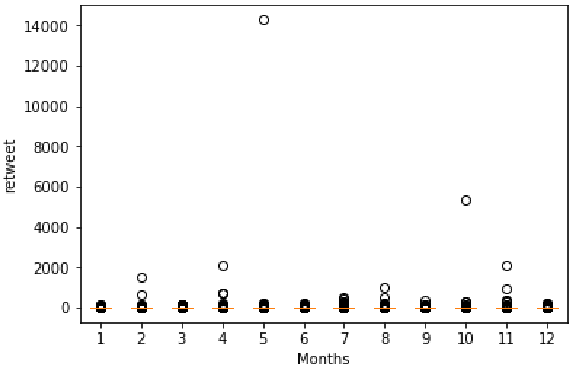


Figure 15: Distribution of retweet on months

Figure 15 shows distribution of retweet on months. It seems that the distribution is close to normal and any specific month does not show the influence on retweet. Hence the retweet regression showed a 0 RSquared, because actually there is no linear relationship exists between months and retweets.

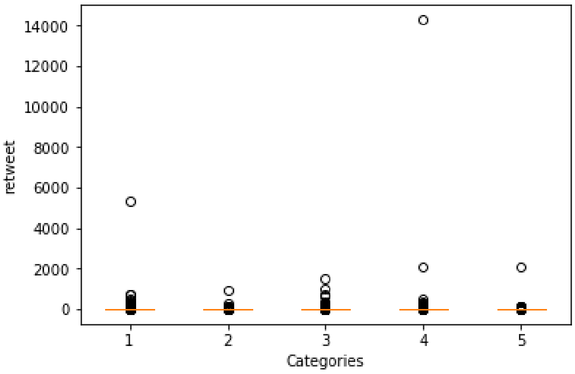


Figure 16: Distribution of retweet on Categories

Figure 16 shows distribution of retweet on categories. It seems that the distribution is close to normal and any specific category does not show the influence on retweet. Hence the retweet regression showed a 0 R-Squared, because actually there is no linear relationship exists between categories and retweets.

We think that to derive any relationship between categories and deal tweets, more research needs to be done to identify the attributes impacting linearly like and retweets. We assumed that categories will have a linear relationship on dependent variables, but as per the above results and analysis it seems there are some other attributes hidden in the tweet text which has a linear relationship on like and retweet. A good example which we have fund was the picture indicator, it showed a linear relationship on likes and retweet. Similarly other attributes can be be found and researched, and then a relation ship between deal categories and attributes(newly found) can be established. In this paper we can find how many categories have picture indicator in tweets. Once we know the categories of tweets having picture indicator as 1, it is easy to establish a linear relationship of these tweets to like and retweet. Similarly more attributes can be found.

Classification approach to predict popular tweets and non popular tweets did not show great accuracy. The accuracy was very near to random baseline, We think the reason for getting this accuracy is the non equivalent number of tweets in each class. The popular tweets are 24644 and non popular tweet are 523183, it means the popular tweets are just 5% of total tweets. Hence algorithm is not getting enough data of both classes to train it perfectly. Another reason which we think for this accuracy is that we classified the tweets based on number of likes and retweets. We did not use tweet text to classify tweets in to popular and non popular tweets. Algorithms are training on the binary version of tweet text, so algorithm is not able to find a relationship between tweet text and like/retweets.Hence generating the accuracy very similar to random baseline. Future studies, as we said above, can be done to get more attributes which has relationship with like and retweet and then it can be used to classify the popular and non popular tweets. Similar to the example of picture indicator, all the tweets can be marked as popular tweets others as non popular tweets.

References

Becker, H., Naaman, M., & Gravano, L. (2011). Beyond trending topics: Real-world event identification on twitter. In *Fifth international aaai conference on weblogs and social media.*

Dayani, R., Chhabra, N., Kadian, T., & Kaushal, R. (2015, Dec). Rumor detection in twitter: An analysis in retrospect. In *2015 ieee international conference on advanced networks and telecommuncations systems (ants)* (p. 1-3). doi: 10.1109/ANTS.2015.7413660

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international acm sigir conference on research and development in information retrieval* (pp. 841–842). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1835449.1835643 doi: 10.1145/1835449.1835643

Theodotou, A., & Stassopoulou, A. (2015). A system for automatic classification of twitter messages into categories. In *Modeling and using context* (pp. 532–537). Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-319-25591-$0_4$4 doi: 10.1007/978-3-319-25591-$0_4$4

Vickey, T. A., Ginis, K. M., & Dabrowski, M. (2013, may). Twitter classification model: the ABC of two million fitness tweets. *Translational Behavioral Medicine*, *3*(3), 304–311. Retrieved from https://doi.org/10.1007/s13142-013-0209-0 doi: 10.1007/s13142-013-0209-0

Wikipedia. (2017, November). *Wikipedia.* Web Page. Retrieved from https://en.wikipedia.org/wiki/Twitter

Yu, B., Chen, M., & Kwok, L. (2011). Toward predicting popularity of social marketing messages. In *Social computing, behavioral-cultural modeling and prediction - 4th international conference, SBP 2011, college park, md, usa, march 29-31, 2011. proceedings* (pp. 317–324). Retrieved from https://doi.org/10.1007/978-3-642-19656-$0_4$4 doi: 10.1007/978-3-642-19656-$0_4$4

APPENDIX

A. TEAM MEMBER CONTRIBUTIONS

All members of the team contributed in collaborating on research question formation, data acquisition plan, method discussion and writing of the project report. Along with these, each of the team members worked on additional responsibilities as mentioned below.

Tanmoy Choudhury Worked on data acquisition, categorization and visualization. Additionally, worked on streamlining codebase, collaboration on research method review and results and discussion and project write up.

Madhavi Pilli Worked on data acquisition, preprocessing, cleansing and research method to gather top tweet words. Identifying and extracting features. Additionally, collaborated on research method and results review , discussion and project report write up.

Vipin Singh Worked on research method and results.Generated various results and analyzed why or why not the models are working.Thought through about implementations and enhancements in future. Additionally, collaborate on research method and results review, discussion and project write up.