Ranking Significant Features for Increasing Engagement on Social Media via Regression Analysis

Thiago R. C. de Lima[1]

[1]Affiliation not available

January 25, 2021

Abstract

Social media comprises of platforms that surpassed their initial goal to connect people just for the sake of socializing and currently provide powerful tools for businesses to reach millions of views worldwide, increasing their chances of gaining new customers. This short paper utilizes the Buzz in Social Media data set available at UCI Machine Learning Repository for identifying the attributes in social media content that have the highest correlation to the amount of repercussion it gained. To achieve such result, several linear regression models are constructed, then ranked based on their respective model fit measure (R-squared) and accuracy when tested against unseen data.

1 Introduction

During the past two decades, the world wide web has seen a great shift on how users interact with the internet. The era of the startups has been a playground for entrepreneurs to try out innovative ways to captivate potential customers and retain users on their platforms and services. While many startups fail or go bankrupt [1] [2], others get sold for millions of dollars [3] and a few find their own success and remain strong. For social media platforms to survive, it is imperative to have an active user base. With the growing registered accounts, and ability to to find out each persons likes and dislikes, such platforms has caught the interest of business wanting to invest on marketing campaigns that have the highest return [4]. The more users interact with each other, the more data can be gathered and a better profile can be set for each user, thus allowing targeted ads to be more and more tailored to each potential customer.

Each social media website has its own algorithm for measuring engagement on specific content provided by some user. Typically such content can be rewarded with increased reach [5]. For instance, on Twitter there is a list of Trending topics. On YouTube, a video may be presented on the home page. Naturally, the more exposed this content gets to users who hadn't seen it yet, the more engagement it might get.

On this research I will analyze data gathered from the Twitter platform. My goal is to find out whether any of the attributes of a tweet has a strong correlation with the amount of discussion it gathered. Some work on this field has been done by [6].

2 Methodology

2.1 Data

The data set used for this research was provided by François Kawala, Ahlame Douzal, Eric Gaussier, and Eustache Diemert (from Université Joseph Fourier and BestofMedia Group) and is currently available at the UCI Machine Learning Repository [7], hosted by the University of California Irvine [8]. This data set contains a total of 40000 rows and up to 96 columns, across data gathered from Twitter and TomsHardware [9]. However, for this research I am using only the Twitter database, which contains 77 attributes for each of its 38393 samples.

Attributes are presented in temporal fashion, varying according to each observation date. Each row contains 7 values each of the following categories: Number of Created Discussions, Author Increase, Attention Level, Burstiness Level, Number of Atomic Containers, Attention Level (measured with number of contributions), Contribution Sparseness, Author Interaction, Number of Authors, Average Discussions Length, Number of Active Discussion. Finally, there is a single value in each row for Mean Number of Active Discussion which I'll use as the target attribute, that is, the one I'm trying to predict.

2.2 Tools

I have developed a script using the Python language and SciPy package. Python is a general purpose programming language [10] [11] that has gained notoriety in the data science field [12] [13] and SciPy is tool set for data analysis, manipulation and visualization [14]. For this particular research, I only used SciPy for calculating the actual linear regression, which returned the slope, the intercept point, the raw R value, the P value and the standard error of the estimated slope [15]. The source code can be found in my repository on GitHub [16].

The script is responsible for loading the data set, converting the data type from strings to floating point representation, partitioning the data set into five folds for cross-validation, then for each predictor attribute, creating a linear regression model and testing such model against the testing portion of the partitioned data. Finally, the script ranks correlations and accuracy of the predictors and picks the ten with the highest scores for each one of the chosen metrics.

For evaluating the effectiveness of an attribute on predicting the target feature, I employed two distinct metrics: first, using the R-squared, also known as coefficient of determination [17], and secondly, the accuracy of models when comparing the yielded result against the known value for the training sample. The accuracy is calculated as the inverted error. The error is given by the difference between the expected value (known value for the target feature in the training data) and the resulting value from feeding the model using the training data as input.

3 Results

In summary, I analyzed a comprehensive data set from Twitter to find attributes that could serve as predictors of the amount of engagement on the comment sections. I applied linear regression models for each feature and cross-validated the results among 5 partitions of data, averaging them and picked the ten most significant features based on two different metrics. The resulting rankings can be found in Table 1 and Table 2.

According to the ranking based on the average R-squared values, the feature with strongest correlation to the sixth observation of Number of Created Discussions, followed by the sixth observation of Number of Active Discussion and Number of Atomic Containers.

On the other hand, when applying the models to the testing data, the highest ranked feature is Number of Authors on its fifth, sixth and fourth observations, respectively. The attributes that generated the most accurate models differ almost significantly. Only four out of the ten best attributes from Table 1 are represented again in Table 2. If we were to pick only the top 5, there would be no recurring attribute at all.

Rank	Attribute	Average R-squared
01	NCD_6	0.91
02	NAD_6	0.91
03	NAC_6	0.91
04	NCD_5	0.85
05	NAD_5	0.85
06	NAC_5	0.84
07	NA_6	0.82
08	NCD_1	0.79
09	NAD_1	0.79
10	NCD_4	0.79

Table 1: Ranking of the ten features highest coefficient of determination

Rank	Attribute	Average Accuracy
01	NA_5	4.91e-06
02	NA_6	9.53e-07
03	NA_4	4.76e-07
04	NAC_2	4.13e-07
05	NCD_3	3.7e-07
06	NAD_3	3.7e-07
07	NCD_6	3.51e-07
08	NAD_6	3.45e-07
09	NAC_3	3.28e-07
10	NCD_5	2.93e-07

Table 2: Ranking of the ten features which produced models with highest accuracy for unseen data

3.1 Observations

The metrics were unable to come up with similar results, that is, even though some features had a strong coefficient of determination, using them to predict the target feature on unseen data resulted in low accuracy.

The inability for a model to perform with unseen data is a typical case of overfitting. For future work, an option would be increasing the amount of folds, or pruning the data set as some temporal data might be affecting the expected outcome.

References

[1] The Venture Capital Secret: 3 Out of 4 Start-Ups Fail. https://www.wsj.com/articles/SB10000872396390443720204578004980476429190. Accessed on Mon, January 25, 2021.

[2] Startup Genome. https://startupgenome.com/reports/global-startup-ecosystem-report-2019, 2019. Accessed on Mon, January 25, 2021.

[3] Scoop: Facebook to buy Giphy for $400million. https://www.axios.com/scoop-facebook-to-buy-giphy-for-400-million-4a75a359-833b-484d-b15b-87e94d3de017.html. AccessedonMon, January25,$ 2021.

[4] M Saravanakumar and T SuganthaLakshmi. Social media marketing. *Life science journal*, 9(4):4444–4451, 2012.

[5] Stefania Milan. When algorithms shape collective action: Social media and the dynamics of cloud protesting. *Social Media+ Society*, 1(2):2056305115622481, 2015.

[6] François Kawala, Ahlame Douzal-Chouakria, Eric Gaussier, and Eustache Dimert. Prédictions d'activité dans les réseaux sociaux en ligne. In *4ième conférence sur les modèles et l'analyse des réseaux : Approches mathématiques et informatiques*, page 16, France, Oct 2013.

[7] UCI Machine Learning Repository: Data Sets. https://archive.ics.uci.edu/ml/datasets.php. Accessed on Mon, January 25, 2021.

[8] UCI Machine Learning Repository — re3data.org. https://www.re3data.org/repository/r3d100010960. Accessed on Mon, January 25, 2021.

[9] Buzz Prediction in Online Social Media — AMA Team. http://ama.liglab.fr/datasets/buzz/. Accessed on Mon, January 25, 2021.

[10] Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[11] What is Python? Powerful, intuitive programming. https://www.infoworld.com/article/3204016/what-is-python-powerful-intuitive-programming.html. Accessed on Mon, January 25, 2021.

[12] Python eats away at R: Top Software for Analytics, Data Science, Machine Learning in 2018: Trends and Analysis - KDnuggets. https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html. Accessed on Mon, January 25, 2021.

[13] Python Developers Survey 2019 Results. https://www.jetbrains.com/lp/python-developers-survey-2019/. Accessed on Mon, January 25, 2021.

[14] Francisco J Blanco-Silva. *Learning SciPy for numerical and scientific computing*. Packt Publishing Ltd, 2013.

[15] scipy.stats.linregress — SciPy v1.6.0 Reference Guide. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.
Accessed on Mon, January 25, 2021.

[16] thiagorcdl/social$_{m}edia_{b}uzz.https://github.com/thiagorcdl/social_{m}edia_{b}uzz.AccessedonMon, January25, 2$

[17] Dabao Zhang. A coefficient of determination for generalized linear models. *The American Statistician*, 71(4):310–316, 2017.