**Predicting the Number of Tweets for an Research Article using Altmetrics Dataset**

**Hari Durga Prasad Karri**

Computer Science, Masters

Northern Illinois University

[Z1830680@students.niu.edu](mailto:Z1830680@students.niu.edu)

**Srikar Akula**

Computer Science, Masters

Northern Illinois University

[Z1816848@students.niu.edu](mailto:Z1816848@students.niu.edu)

**Rahul Marupaka**

Computer Science, Masters

Northern Illinois University

[Z1802041@students.niu.edu](mailto:Z1802041@students.niu.edu)

**Abhi Sekhar Reddy Dwarampudi**

Computer Science, Masters

Northern Illinois University

[Z1816848@students.niu.edu](mailto:Z1816848@students.niu.edu)

**ABSTRACT**

In simple definition, in classification/clustering analyze a set of data and generate a set of grouping rules which can be used to classify future data and this paper presents the overview of machine learning techniques in classification of Tweet counts using Random Forest, Naive Bayes and Decision Tree Classification methods. As mentioned, Classifying the future data here talks about predicting the data of a certain feature in the future. As part of regression techniques for modeling and analysis of numerical data consisting of values of a dependent variable and of one or more independent variables, we have considered and performed Linear, Random Forest and Lasso regressions on our data in order to come up with the best model for the prediction.

**Keywords**

Classification, Regression, Clustering, Regression, Modeling

# INTRODUCTION

In this project, we tried to predict the number of tweet counts for a certain article using the altmetrics data. Citation has been the widely accepted metrics of impact of a scientific publication for decades. In the recent years, many indicators have been introduced to the family of article-level metrics. Besides citations, the impact of an article could be reflected and quantified by article views, readerships and Altmetric score, etc. Altmetrcis tracks the online mentions by pulling in data from social media, blog, traditional media and online reference managers. Recent efforts have been made to investigate the effect of altmetrcis on scholarly impact.

**What are altmetrics?**

Altmetrics are a broad class of statistics which attempt to capture research impact through non-traditional means. In simple terms, they are ‘new metrics based on the social web for analyzing, and informing scholarship’. The sources mined for altmetric data include:

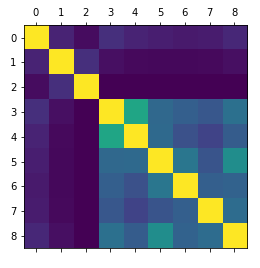
* micro-blogging or short-message services (Twitter),
* social networking sites (Facebook),
* blogs (WordPress, Blogger),
* social bookmarking networks (Delicious),
* academic bookmarking platforms (CiteULike, Mendeley),
* peer review services (F1000, now F1000Prime),
* academic networks (Academia.edu),
* collaboratively edited online encyclopaedias (Wikipedia)

Data from these sources are potentially subject to multiple forms of analysis. As mentioned earlier, we used the altmetrics data for this project. There are approximately 250k rows of data that has been extracted from Altmetrics JSON files. We converted the JSON files in to the Comma Separated Value documents.

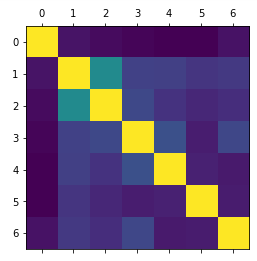
For this project, we first started with the data exploration. This is the first step in data analysis and it typically involves summarizing the main characteristics of a dataset. It is commonly conducted using visual analytics tools but can also be done in more advanced statistical software. The next step is data cleaning. It is the process of cleaning / standardizing the data to make it ready for analysis. Most of times, there will be inconsistencies in the captured data such as incorrect data formats, missing data, errors while capturing the data. This is a crucial step in any given data science project because the accuracy of the results depends heavily on the data we use. The next step is training and testing the data. If you evaluate your model on the same data you used to train it, your model could be very overfit, and you wouldn’t even know! A model should be judged on its ability to predict new, unseen data. Therefore, you should have separate training and test subsets of your dataset. Training sets are used to fit and tune your models. Test sets are put aside as "unseen" data to evaluate your models. You should always split your data before doing anything else. This is the best way to get reliable estimates of your models’ performance. After splitting your data, don’t touch your test set until you’re ready to choose your final model! For our project, we used 80 percent of the complete data as training data and the rest of the 20 percent of the data as test data.

# Feature extraction & Elimination:

We have extracted the Data from altmetrics dataset into JSON files using java. This provided us nearly 3,00,000 records. The data has been converted into csv files. Out of which we have a lot of features consisting of 0’s. Since they are not very useful for the prediction process we have get rid of these records with 0’s in the features that are considered. The features are chosen to be eliminated based on the 0’s that are present in the available data and their contribution towards the prediction. If the features that are considered has a 0 in it, the whole record has been discarded in our experiment. The features are selected using **Heat Map**. So, using the data the Heat Map has been analyzed. The low correlated features based on the HeatMap coloring technique are considered to be the best features for “**Predicting the Number Of Tweet Counts for an Article**”



**Fig 2.1. Heat Map of all the considered features**



**Fig 2.2. Heat Map after Reduced Features**

**2.1 Final Features:**

mendeley\_count,blogs\_count,news\_count,facebook\_count,googleplus\_count,reddit\_count,citeulike\_count, connotea\_count are the final features with low correlation. These are to be considered in order to get the better prediction. Outliers from the data has been removed i.e., since the outliers has a great impact in the prediction, they are to be removed. In our experiment the records with twitter count more than 350 is removed.

|  |  |
| --- | --- |
| **Features** | **Correlation Value** |
| Mendeley\_count | 0.111057968093 |
| citeulike\_count | 0.040933519646 |
| connotea\_count | -0.00082188390 |
| blogs\_count | 0.369947651581 |
| news\_count | 0.288494654766 |
| facebook\_count | 0.487900992762 |
| googleplus\_count | 0.312678284114 |
| reddit\_count | 0.353231360281 |

**Table 1.1 Correlation Coefficient of Tweet\_counts with other features**

# Regression Analysis:

Regression is a statistical measure used in finance, investing and other disciplines that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables. It is Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships. The parameters are estimated so as to give a "best fit" of the data.

## Liner Regression:

linear regression is a linear approach for modelling the relationship between a scalar dependent variable *y* and one or more explanatory variables (or independent variables) denoted *X*.

In our project we used multivariate linear regression. Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine

## Random Forest Regression:

Random Forest Regression, is an ensemble learning for regression that operate by constructing a multitude of decision trees at training time and outputting the mean prediction.

A random forest Regression is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.

## Lasso Regression:

Lasso (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.

Though originally defined for least squares, lasso regularization is easily extended to a wide variety of statistical model including generalized linear models, generalized estimating equations, proportions hazards models and M–estimators, in a straightforward fashion.

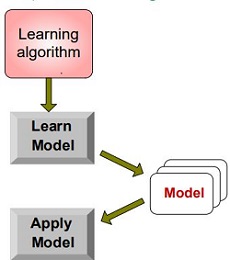
|  |  |  |
| --- | --- | --- |
| **Regression Type** | **Mean Square Error** | **R2 score** |
| **Linear** | 0.000 | 1.000 |
| **Random Forest** | 6.855 | 0.998 |
| **Lasso** | 2.038 | 0.999 |

**Table: 1.2. Results for Regressions**

From the Table1.2, we can see the results what we got for regression models. In that the Linear regression is performing accurately by projecting mean square error and r-square with 0 and 1 respectively and Random forest and Lasso regression is also projecting mean square error and r-square which is high than Linear Regression. So by doing Regression, We have finalized that Linear Regression is best Regression model for predicting the tweet counts perfectly and we extended our work to check whether tweet counts is high or low using classification models.

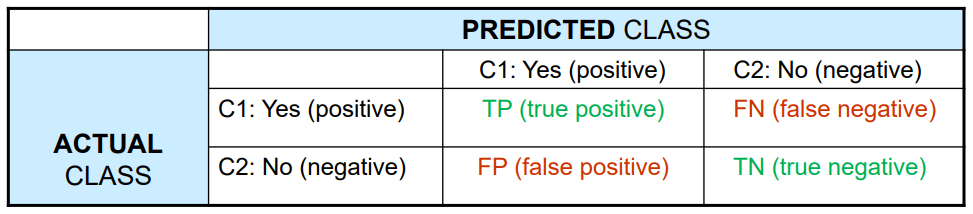
# Classification

In order to check the models compatibility, The data has been classified into two class model. From collection of records i.e., Training set each record contains a set of attributes (features), one of the attributes is the class label (output, target, category). Classification is the task of learning a target function f (classification model) that maps each attribute set x to one of the predefined class labels y. It is most suited for describing or predicting datasets with binary or nominal categories and we have chosen the features having numerical data. In general, Classification has two step process which we have applied this two-step process on dataset. Model Construction is the first step which will describe a set of predetermined classes and Second step is model Usage for classifying future or unknown objects.



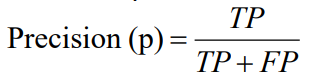
**4.1 Model Construction:**

The features that have been selected is classified as either Low or High class by taking the Mean(Average) value of tweet counts and now each tuple is assumed to belong to a predefined class, as determined by the class label attribute with low or high label. Now the dataset is divided into training and testing data by taking train and test percentage with 80% and 20% respectively with random state of random function. While doing the test and train split sometimes we got low accuracy rate because the model was not able to learn the data and apply the model in predicted way for that we have used the random state with 42 which gave us the high accuracy rate every time. After training the data we have applied the three classification methods like Decision Tree, Random Forest and Naïve Bayes Algorithm. The Classification methods will take the 80% train data(target and Non-target columns) and learns the model and create a model based on the train data. Later it will apply the model on test data like follows the above figure and produce the accuracy rate. By using accuracy rate we can determine which classification methods is performing better. precision and recall are used to know how the data is predicted. The below formulae are precision and recall and the perfect score is having value close to 1.0 and Precision and recall scores are typically used together, where precision values are compared for a fixed value of recall, or vice versa. A system with high recall but low precision returns many results, but most of its predicted labels are incorrect when compared to the training labels. A system with high precision but low recall is just the opposite, returning very few results, but most of its predicted labels are correct when compared to the training labels. An ideal system with high precision and high recall will return many results, with all results labeled correctly.



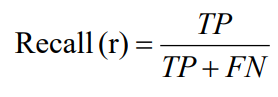
**Precision:** exactness:

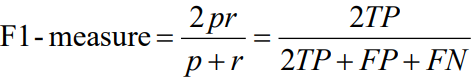
what % of tuples that the classifier labeled as positive are actually positive ?



**Recall:** completeness:

what % of positive tuples did the classifier label as positive ?





**Accuracy** = (TP + TN) / ALL

**4.2 Model Usage :**

Model Usage is for classifying the future or unknown objects.The test data with classified label feature is get predicted using the classification models and generates the accuracy rate for each classification model. The known label of test sample is compared with the classified result from the model to cross check where the predicted value is overfitting or getting underfitting. Before applying Classification models we have reduced our dataset by removing the outliers of tweet counts feature to get better accuracy rate and less error rate.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Name** | **Precision** | **Recall** | **F1 score** | **Accuracy** |
| **Decision Tree** | 0.81 | 0.83 | 0.80 | 0.830 |
| **Random Forest** | 0.82 | 0.83 | 0.80 | 0.832 |
| **Naïve Bayes** | 0.80 | 0.82 | 0.79 | 0.823 |

**Table** **1.3 precision, recall, F1 score and Accuracy values for the Models used**.

The Table 1.3 shows the prediction for tweeter count of Altmetrics dataset. From the Classification methods, we got precision,recall,F1 score and Accuracy for Test dataset and We got High precision and recall values for Random Forest Classification when comparing with other classification Model. The Random Forest Classification is having Highest Accuracy rate with 0.823 then Decision Tree and Naïve Bayes. We can see the Accuracy rate for Decision Tree and Naïve Bayes is bit close to the Random Forest but by considering the precision and recall Values we have decide that Random Forest is predict the test data with less error rate. Finally, we have predicted where the tweet counts for the given feature data will be high or low using classification methods.

# CONCLUSION AND FUTURE WORK:

In conclusion, from the Altmetrics dataset we have extracted the required features and applied the classification methods on the data features. From the classification methods, we have predicted where the tweet counts will be high or low class label and finalized that Random Forest Classification is performing better than other classification methods. For future work, we can consider the feature like date of tweet to predict the raise of tweets for an article.

# REFERENCES

1. Priem, J., & Costello, K.L. (2010). How and why scholars cite on Twitter. In Proceedings of the 73th Annual Meeting of the American Society for information Science and Technology, Pittsburgh, USA. doi: 10.1002/meet.14504701201
2. Priem, J., Costello, K.L., & Dzuba, T. (2011). Prevalence and use of Twitter among scholars. Poster presented at Metrics 2012: Symposium on Informetric and Scientometric Research, New Orleans, USA. doi: 10.6084/m9.figshare.104629
3. Priem, Jason, Paul Groth, Dario Taraborelli, and Christos A. Ouzounis. 2012a. “The Altmetrics Collection.” PLoS ONE 7 (11): e48753. doi:10.1371/journal.pone.0048753.
4. Priem, Jason, Heather Piwowar, and Bradley M. Hemminger. 2012b. “Altmetrics in the Wild: Using Social Media to Explore Scholarly Impact.” Accessed May 16, 2014. <http://arxiv.org/abs/1203.4745v1>
5. Priem, Jason, Dario Taraborelli, Paul Groth, and Cameron Neylon. 2010. “Altmetrics: A Manifesto.” Accessed March 21, 2014. <http://altmetrics.org/manifesto/>
6. K. Alsabti, S. Ranka, and V. Singh. CLOUDS: A Decision Tree Classiﬁer for Large Datasets. In *Proc. of the 4th Intl. Conf. on Knowledge Discovery and Data Mining*, pages 2–8, New York, NY, August 1998.
7. L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone. *Classiﬁcation and Regression Trees*. Chapman & Hall, New York, 1984.
8. W. Buntine. Learning classiﬁcation trees. In *Artiﬁcial Intelligence Frontiers in Statis- tics*, pages 182–201. Chapman & Hall, London, 1993.
9. V. Cherkassky and F. Mulier. *Learning from Data: Concepts, Theory, and Methods*. Wiley Interscience, 1998.
10. K. Lee,Jalal Mahmud, Jilin Chen, Michelle Zhou, Jeffrey Nichols "Who Will Retweet This? Automatically Identifying and Engaging Strangers on Twitter to Spread Information".
11. Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J.Everyone's an influencer: quantifying influence on twitter, In *WSDM,* 2011.
12. Starbird, K. and Palen, L. Pass It On?: Retweeting in Mass Emergency, In *ISCRAM*, 2010.
13. Yarkoni, Tal. Personality in 100,000 words: A large-scale analysis of personality and word usage among bloggers. *Journal* *of Research in Personality,* 2010.
14. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. The WEKA data mining software: an update. *SIGKDD Explorations Newsletter*, 11(1): 10-18, 2009.
15. Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett.*, Vol 27( 8), 2006.
16. Ver Steeg, G. and Galstyan, A. Information transfer in social media. In *WWW*, 2012.
17. Fast, L. A., and Funder, D. C. Personality as manifest in word use: Correlations with self-report, acquaintance report , and behavior. *Journal of Personality and Social Psychology*,Vol 94(2), 2008.