Which articles receive much attention on social media?

Sai Harini Perugupalli
Z1829025
Northern Illinois University
Email: harini496@gmail.com

Krithin Kumar Venkatesh
Z1837607
Northern Illinois University
Email: venkateshk284@gmail.com

***Abstract*— Social media can be defined as "we-based communication tool" which enables users to share and consume information. The usage of social media has been increased since last few years. In other words, social media usage has affected several fields of study like politics, healthcare, education and research. Social media is the best platform to post about one's research. In this paper, we demonstrate how can we predict the popularity of an article based on the Facebook social media content. Based on the number of likes, which is a target variable in the dataset we predict the highest attention gained article. We performed linear regression and neural network algorithms on the dataset and drawn a conclusion which type of articles are being popular.**

I.INTRODUCTION

Social media gave people an opportunity to be content creators, controllers and transparent users, to a great extent. In other words, social media is an act of engagement where users can share their point of view on miscellaneous topics. Because of its ease of use, speed and reach, social media is fast changing the public discourse in society and setting trends and agendas in topics that range from the agriculture to research industry. We have numerous websites for showcasing your work, discovering research data and collaborating online. Examples include Facebook, Mendeley, citeulike, blogs, news, Wikipedia, google plus, qna, reddit and altimetric. Since social media can also be construed as a form of collective wisdom, we decided to predict real-world outcomes on social media data. Our paper reports one such study. Surprisingly, we discovered that the chatter of a community can indeed be used to make quantitative predictions that outperform those of artificial markets.

The one's who get benefitted from our study are:

- Organizations which buy copyrights from the authors based on their popularity.
- Students who can use the sources of high ranked authors, it strengthens their research paper as well.
- Can know what field people are really interested in

We considered the task of predicting the popularity of an article in social media. In other words, which type of paper gets more citation counts in future. We used the altimetric dataset provided by NIU where we had millions of JSON files and each JSON file represented a tuple in the altimetric data, which were processed to create the final dataset. Due to the large dataset, we extracted around 3,00,000 tuples using random. The features in our dataset are Mendeley_count, citeulike_count, connotea_count, blogs_count, news_count, Wikipedia_count, facebook_count, googleplus_count, qna_count, policy_count, reddit_count and altmetric_score. Secondly, we cleaned the dataset using the various cleaning techniques which will be discussed later in the below section. We then converted the JSON file to a csv file and then separated the target variable, facebook_count and considered rest of the data as test data (i;e drop the facebook_count feature and considering all other features). The reason why we considered the facebook_count as target variable is because it is one of the largest social media platforms.

Our goals in this paper are as follows. First, we split the data using train_test_split from sklearn and then performed linear regression algorithm which will be explained clearly in below sections. Next, we scattered a plot to know how similar the features in the given dataset are. Later, mean square error is calculated to know how close a fitted line is to data points. Then, we performed the above steps considering altmetric_score as a target variable and rest of the data as test data.

Next, we performed neural networks on the data where the target variable is facebook_count and calculated mean square error. We observed that neural networks performed better when compared to the linear regression.

This paper is organized as follows. Next, we survey recent related work. We then provide a short introduction to the dataset that we collected. We then discuss our study using algorithms, linear regression and neural networks. We conclude in the last Section.

I.RELATED WORK

. Tarek A. El-Badawy and Yasmin Hashem [1] studied the impact of social media on the academic development on students. They conducted chi square analysis between the use of social media and number of hours spent on studying. There is no significant relationship between using social media and the students' academic performance.

Xianwen wang, Yunxue Cui and others [2] studied the social media attention increases article visits. We have gone through

many papers related to our project, but these were the citations which are most related to our project. These papers have retrieved their data from plum analysis which has been integrated with into Scopus. They have performed correlation analysis between Facebook and twitter scores to find out the social media attention.

Gemma Nandez and Angel Borrego [3] studied the use of social networks for academic purposes. The results are based on a single case study. This study provides new insights on the impact of social media in academic contexts by analyzing the user profiles and benefits of a social network service that is specifically targeted at the academic community.

Stefanie Hasten and others [4] studied characterizing social media metrics of scholarly papers. Social and mainstream media metrics analyzed in this paper include scientific blogs, Twitter, Facebook, Google+ and mainstream media and newspaper mentions, as covered by Altmetric.com. By combining these various social media sources with traditional bibliometric indicators, this paper aims to perform the first large-scale characterization of the drivers of social media metrics and to contrast them with the patterns observed for citations.

II.DATASET CHARACTERISTICS

The altimetric dataset of size 27GB was obtained from NIU Big data course. The data was in Json (JavaScript-object notation) file format. There were several millions of json files where each json file represented a tuple in the altimetric data. The json file format was difficult to be processed and analyze directly. Hence, we converted the data in json file format into a CSV (comma-separated values) file format and extracted the dataset of 12 columns where each column represents the attributes of the dataset and 235771 rows where each row represents a tuple of respective values of specific attribute. Due to the large dataset, data cleaning seemed to be a challenging work for us. We used the following cleaning techniques to clean the data:

- Remove outliers
- Dropna() to remove null values using numpy
- Imputer to fill in the missing values using sklearn
- Drop_duplicates() to remove duplicate values

Each attribute in the dataset is a specific count of each paper on various social media platforms. The considered social media platforms are Mendeley, citeulike, connotea, news, Wikipedia, Facebook, google plus, qna, policy, reddit and altimetric. Our main goal was to analyze the data from eleven attributes and make a prediction of popularity on a target variable, facebook_count. Since, Facebook has a greater number of active users, we chose facebook_count as a target variable. The features were explained in Table 1.

Attribute	Social media Platform
mendeley_count	The number of Mendeley users that have added to a particular document to a Mendeley library
citeulike_count	An examination of citation counts in new scholarly communication environment
connotea_count	Number of views a paper gets in connotea org
blogs_count	Count indicating the number of times a web page has been loaded.
news_count	Number of times a paper viewed in news
wikipedia_count	The total times the user have viewed a paper in Wikipedia
facebook_count	number of unique users who saw your Page post in News Feed, Facebook
googleplus_count	counts for any view of them in any Google+ stream
qna_count	Counts on qna platform
policy_count	Counts a paper gets on Policy
reddit_count	Counts a paper gets on Reddit
altmetric_score	attention that research outputs such as scholarly articles and datasets receive online.

Table 1
Attributes and what they mean

We then plotted graphs for the features' vs facebook_count (number of likes) to know how similar they are. Later, we plotted histograms for the above listed features and calculated the Pearson coefficient. Based on our calculations, the Pearson coefficient for connotea_count and facebook_count is very low. Then, we considered few datapoints (50, 100, 150, 200) and generated a heatmap for the distance matrix to know which features are similar. In later section, we explain about the algorithms performed on the dataset in detail.

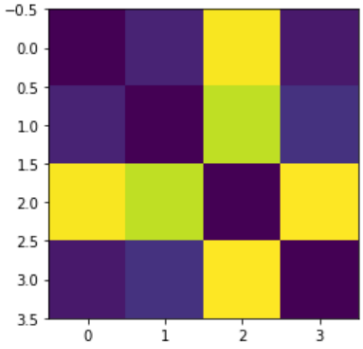

Fig1: Heatmap

III.ALGORITHMS

The algorithms we used to process, analyze and predict the data are the Linear regression and Artificial Neural networks. Since our data is continuous, we performed linear regression. Our dataset was huge, so we didn't get higher rate of accuracy. So, that's the reason why we conducted neural networks algorithm on our dataset. Our results and calculations were explained in detail in the later sections.

LINEAR REGRESSION

Linear regression is one of the most efficient regression an alysis algorithms which can be used to fit a model which is pre dictive to an observed data. It is the most commonly used type of predictive analysis. These regression estimates are used to explain the relationship between one dependent variable and o ne or more independent variables.

So, we split the dataset into test and train data sets before w e could proceed with the regression analysis where Facebook a s the target variable. We imported the linear regression algorit hms from sklearn library. This process is mainly used because it can be used to model the relationships from unknown model parameters through linear functions. The linear regression is d one with analysis done on the Test set and predicting the result ant target set for the respective test set. When we implemented the Linear regression on the gained dataset, we received fair p redictions.

These predictions were measured for efficiency which was done using one of the ways to evaluate linear regression. We o pted for mean square error(mse). The root mean square error [5] represents the sample standard deviation of the differences between predicted values and observed values. It is calculated by using the above formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (P_i - O_i)^2}{n}}$$

The MSE was around 3.20. This indicated that the predictio ns received were fair with less errors or variations on the result ing predictions.

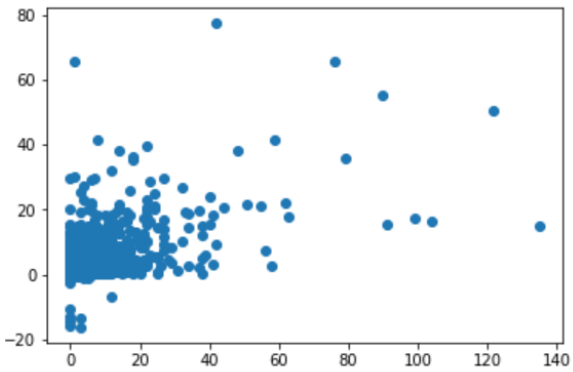

Fig2: Scatter plot with Facebook as Target.

We also implemented the Linear regression on the dataset with Altimetric score as the Target variable just to check which dataset would give the best Linear regression predictions.

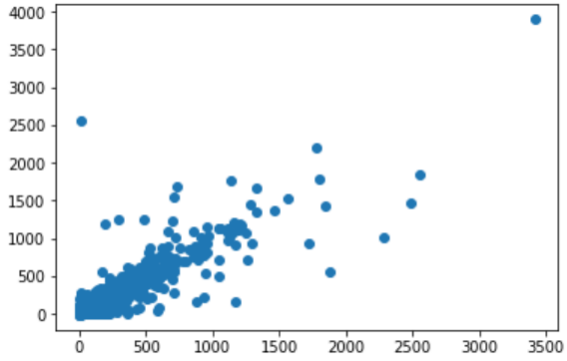

Fig3: Scatter plot with Altimetric score as Target.

NEURAL NETWORKS

A neural network is a series of algorithms that helps us to srecognize the underlying relationships in a dataset through a process, just like the human brain. Though it is complexed, it can adapt to a changing input so that the network generates the best outcome without redesigning the output criteria.

We implemented the neural networks on the data frame from keras [6] where the target variable is facebook_count. In deep learning, epoch [7] is a hyperparameter defined before training a model. One epoch is when an entire dataset is passed both forward and backward through the neural networks once. We considered 50 epochs, which means the model will be trained 50 times. In this process, the loss/error will be considerably lowered.

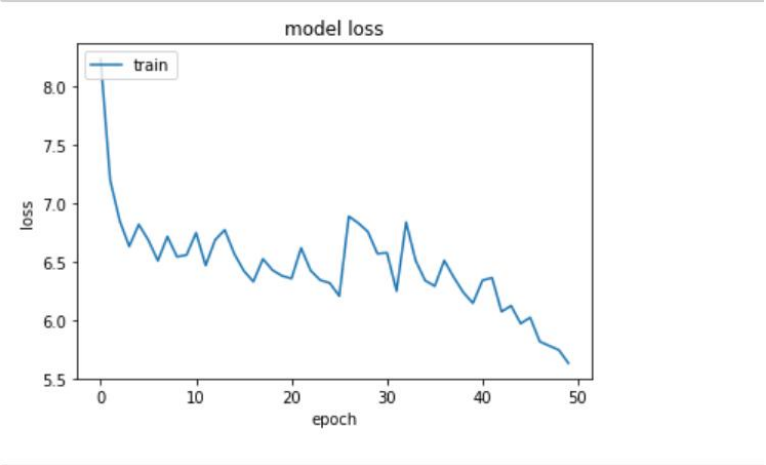

Fig 3. Plot between epoch and loss

When we measured the mean square error for evaluating the neural networks, our error rate came down to 2.0 from 3.2. Thus, neural networks was an improvement in predicting the values with reduced error rate.

IV.CONCLUSION

In this article, we have focused on how social media platforms can be utilized to predict the popularity of the paper in a specific focused social media platform. Using a data from given altimetric data from NIU, we have cleaned the data and converted the data in JSON file format into a csv format to do efficient processing and analysis of the data. Further, we implemented the Linear regression model on the data to predict the required values. The outcome of this was pretty convincing as the predictions came out with good accuracy. But this model was outperformed by Artificial neural network model which was performed with Facebook as the Target variable and predicting the values in that specific Platform. The outcome of this model has outperformed the Previous model with a mean square error reduced considerably compared to the previous one.

We mainly focused on analyzing the data more efficiently using the data and giving out more accurate predictions using the linear Regression and Artificial neural network models. One of the problems we faced in this study was giving efficient predictions using the semi-structured data which was overcome using the python language code used to convert the data in desired format. Further methods can be implemented on this dataset to verify and check for the accuracy of the predictions. This paper describes how many social media platforms can be used as basis to make predictions to know the popularities of papers in several other platforms.

V.REFERENCES

[1] https://www.researchgate.net/publication/273770861_The_Impact_of_Social_Media_on_the_Academic_Development_of_School_Students

[2] https://arxiv.org/ftp/arxiv/papers/1801/1801.02383.pdf

[3] https://www.researchgate.net/publication/263612647_Use_of_social_networks_for_academic_purposes_A_case_study

[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4363625/

[5] http://statweb.stanford.edu/~susan/courses/s60/split/node60.html

[6] https://machinelearningmastery.com/regression-tutorial-keras-deep-learning-library-python/

[7] http://www.fon.hum.uva.nl/praat/manual/epoch.html

[8] https://dl.acm.org/citation.cfm?id=1914092

[9] https://dl.acm.org/citation.cfm?id=1963309.