Which articles will receive much attention on social media sites?

Name: KRITHIN KUMAR VENKATESH (Z1837607)
SAI HARINI PERUGUPALLI (Z1829025)

Introduction:

Nowadays social media plays a critical tool in scholarly papers. According to the social media facts, the networks with the most penetration among social media users in 2018 so far are Facebook, Instagram and Snapchat. In this paper, we analyze which article receives much attention on the social media platform: Facebook. The features which are considered in this paper are: author rank, number of papers published by the author, number of citations for the author, the team size, the publication venue, reference count, number of fields being cited by a paper. The most challenging work in our study is data extraction. The data extracted from online sources is in the ". json" format. Added to that, we have numerous .json files i: e millions of data. So, in order to avoid this problem, we have selected around 70,000 data fields. Algorithms, Linear regression and neural networks are used to predict the most attention gained paper.

Problem significance: Why should we care? What is the need? Who will benefit?

When constructing a research paper, it is important to include reliable sources. Academic research papers are typically based on scholarly sources and primary sources. When using the sources of high ranked authors, it strengthens your research paper. The ones who get benefitted from this paper are the organizations, authors, students.

Research hypotheses: What questions are you trying to answer?

If a paper is published, we can predict how much attention the paper will gain in a social media platform which will be based on the above-mentioned features.

Related work: What other work has been done before? Make sure that you cite appropriate related work and provide a list of references at the conclusion of your proposal (at least 20 citations).

https://arxiv.org/ftp/arxiv/papers/1801/1801.02383.pdf

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 attention whereas in our project we have focused only on Facebook to show the social attention.

https://www.researchgate.net/publication/263612647_Use_of_social_networks_for_academic_purposes_A_case_study

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. Whereas our project is focused on articles popularity on social media.

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

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.

Data:
The dataset used for this project will be the altimetric dataset provided to us in the big data course at NIU. This dataset consists of articles and citations, which consists of features including: author rank, number of papers published by the author, number of citations for the author, the team size, the publication venue, reference count, number of fields being cited by a paper.

Methods:

Linear Regression:

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. Linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Linear regression is used for continuous data in our project. We are going to use scikit-learn, TensorFlow, Theano and keras machine learning library for our project.

Artificial neural networks:

Neural Networks is a framework for many machine learning algorithms to work together and process complex data inputs. Such systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules. In image recognition, they might learn to identify images that contain cats by analyzing example images that have been manually. They automatically generate identifying characteristics from the learning material

that they process. In our project, we are going to use the artificial neural networks is initially trained or fed large amounts of data. Training consists of providing input and telling the network what the output should be.

Innovation: What differentiates your proposal from earlier work? How is it new?

In our project, we are implementing the Artificial neural networks along with the Linear regression for better analysis (data evaluation, processing and predicting the Target values).

Evaluation: How will you evaluate your project? What metrics will validate your results?

In our Project, we are going to evaluate the metrics with many different methods such as Confusion matrix, AUC-ROC, Gini co-efficient, Root mean squared error and cross validation. These are done to make our model efficient with reduced error and more accurate values at output.

Time plan:
11/17/2018: Collect and clean up raw data
11/22/2018: Buildup data training model
11/28/2018: Refine data training model
12/4/2018: Visualize result and prepare for the presentation
12/8/2018: Final presentation

Expected results:

We will be using many features straight from the dataset and also features that we have created/generated to get a better and more precise result. Our research will potentially use the Artificial neural networks. There are many different ways to get a prediction model but based on our dataset, we believe that this is the best way to approach this problem. Build a prediction model using the above methods to obtain the maximum efficiency in each model by feature analysis and create a visualization to represent the results and differentiate the results between training model and test model.