Predicting the Presence of Scholarly Articles in Twitter

Miftahul Jannat Mokarrama
Department of Computer Science
Northern Illinois University
De Kalb, Illinois
z1936043@students.niu.edu

Bhavana Ramineni
Department of Computer Science
Northern Illinois University
De Kalb, Illinois
z1904418@students.niu.edu

ABSTRACT

With the growing presence of research shared on social media; it is important to predict the scholarly articles that are getting more shares on Twitter. The need for the prediction is to measure the social impact of and public attention to scholarly articles by analyzing buzz in social media. Predicting the number of shares of articles has practical significance in evaluating the influence and the value of the article. Our project goal is to predict the presence of a scholarly article in Twitter from the significance of the attributes that are promising to drive the user-attention. The proposed approach creates a standard Twitter dataset from the existing noisy original one from *altmetric.com* that can be used in future for more advanced and diverse analysis. In this approach, we perform preprocessing to remove duplicate entries, check for null values and noisy entries and extract unique values. We used SVM, KNN, and Random Forest classifiers to predict the number of shares and found that KNN performed best with an accuracy of 81.98%, f1-score of 81.51%, precision of 77.2%, and recall of 72%.

KEYWORDS

Machine learning algorithms, Prediction, Scholarly articles, Twitter

Introduction

In recent years social media, particularly, Twitter has become part of scientific communication, as researchers increasingly turn to social media platforms to discuss scholarly articles by posting them. With a service like Twitter, scientists can readily reach audiences allowing their ideas and opinions to diffuse [1]. Tweet shares should be primarily seen as a metric for how quickly new knowledge is taken up by the public as well as a metric to measure public interest in a specific topic [2]. It is also important for understanding and controlling information diffusion on Twitter. Contemporary methods usually consider it as a classification or regression problem. In this project, we predicted

the expected twitter shares for articles using multi-class classification from important features in the data. Our goals are:

1. Create a standard Twitter dataset from the existing noisy one that can be used in future for more advanced and diverse analysis.
2. To find out the promising classifiers to predict the presence of scholarly articles in Twitter
3. Predict the presence of a scholarly article in Twitter given the attributes.

There are a few differences between the work we did on this project and the previous works. In this project, we are using a dataset from altmetric.com which consists of noise and inconsistency. Therefore, the first innovation is the reformation of the dataset. The second innovation is that we worked on scholarly articles which have very less public attention when compared to sports, news, and movies etc. All the research papers are mainly on popularity and information diffusion in social media. There are no research papers on the prediction of shares of scholarly articles. We introduced this in our project to get the user's attention to scholarly research articles on Twitter. We used multiclass classification and got the basic classifier outputs. This will be very useful for the researchers in their future work. They will also get an idea of what approach they should use to improve accuracy and what methods should be applied on the dataset by further feature based modification.

Related Works

Information diffusion is a process in which a new idea or an action widely spreads through communication channels [3]. Jansen and others [4] have examined Twitter as a mechanism for word-of-mouth advertising and considered brands and products while examining the structure of the postings and the change in sentiments. However, the authors do not perform any analysis on the predictive aspect of Twitter. We treat retweet trees as communication channels of information diffusion and observe that retweets reach a large audience and spread fast. Boyd et al. [5] took a detailed analysis of retweets in Twitter to explore how and why people retweet a tweet. Bandari et al. [6] proposed a model to predict popularity of new articles.

They classified articles into three classes based on their retweet number 1-20, 20-100 and 100-2400. The model can predict ranges of popularity on Twitter with an overall 84%

accuracy. Most of the studies above focus on Twitter and few studies are on direct retweet numbers. There are also a few works on retweet number prediction of Chinese micro-blog. Li Ying-le et al. [7] proposed a prediction model based on SVM algorithm with five features: user influence, user activity, interest similarity, the importance of micro-blog content and users' closeness.

Zhang Yang et al. [8] analyzed the importance of different features and investigated the feasibility of applying classification methods and proposed a feature-weighted model. Their model can predict a major fraction of tweets (nearly 86%). Hong, et al., [9] predict whether or not a tweet will be retweeted and how many times a tweet will be retweeted. They employ logistic regression in their work. How many times a new tweet will be retweeted based on a certain threshold is studied by Jenders, et al., [10]. They utilize models such as Naive-Bayes and generalized linear models. Gao, et al.,[11] also predict how many times a tweet will be retweeted by applying an extended reinforced Poisson process model with a time mapping process. Xu, et al., [12] analyze user retweet behavior at individual level and argue that the most important features for general people are social. Morchid, et al., [13] study the behavior of tweets that have been massively retweeted in a short time. Specifically, they employ Principal Component Analysis to select features. Compared to our work, they extract a smaller number of features and number of learning methods.

Suh et al. [14] explored the relationship between different features of tweets and their shares. They showed a strong positive correlation of retweet rate with number of followers of the author and age of the Twitter account of the author. Chen et al. [15] used different features based on the similarity between author and follower to develop a tweet recommendation model. Kwa et al. studied the relationship between the number of followers of the authors of the tweets and their retweets for a collection of 106 million tweets. The authors constructed retweet trees and examined tree temporal and spatial characteristics. Remy et al. studied the impact of the number of followers of users on the capacity to propagate their message. Interestingly, they showed that the impact of users with a lot of followers is not statistically greater than users with a few followers [16]. Kwak and Lee (2014) studied scholarly sharing in Twitter communities and found that measuring scientific impact through the lens of social media to be of limited value given the focus on just a few top journals [17]. Zhu, Turney, Lemire, and Vellino (2015) proposed a machine learning approach using various features to identify the cited articles with the greatest influence on a given publication. They found that the number of times

an article is cited in the body of the article is the most important feature for this classification task [18]. Peng et al. 2011 studied retweet network propagation trends using conditional random fields, demonstrating gains in accuracy when considering social relationships and retweet history [19]. Hansen et al. 2011 investigated the features of tweets contributing to shares and was the first to explore the impact of negative sentiment in diffusion of news on Twitter [20]. Hoang and Mothe (2018) proposed a predictive model for information diffusion on twitter based on user-based, content-based, and time-based features to evaluate if a post will be retweeted or not [21]. Cranmer [11] also used regression modeling to predict clinical trial outcomes. They used machine learning practices to validate their model using data that the models had not yet seen (i.e., testing data), showing how their models performed in uncertain circumstances; they found, however, that though their models might fit well on the data used for training, the results on new data were not reliable.

Dataset Description

The dataset we used in our project is collected from *altmetric.com*. There were more than 10MB entries with a huge number of duplicates and noise that make the file size of around 13.5 GB. We have removed the duplicates and got the new file with 4,35,843 unique entries with the file size of around 86.5 MB. The data we used have a total of 26 attributes. Among them we worked on 10 attributes. Following is the short overview of the attributes that we used in our project. The dataset after removing duplicates is available in this: Twitter_data

Methods

To achieve our goal, we worked in three steps:

1. *Preprocessing:* If we investigate the sample of the data for each attribute in Table 1, it is obvious that they are not in an appropriate format to start working on instantly. So, at first step our goal was to make a standard dataset on which we can do the classification as classification algorithms are susceptible to noise and change their performances based on the quality of the input data. we removed duplicate entries, checked for null values and noisy entries in different stage based on different criteria, extracted the unique values from each attribute type. In Table 2, some preprocessing samples are shown.

2. *Data Filtering:* In Table 1 it is also obvious that 8 out of 10 attributes are categorical or nominal types. So, we converted those categorical values to binary values by encoding. We used *binary encoding* by manual coding logic and *one-hot encoding* techniques for this purpose where appropriate. An obvious problem that arises in this process is the curse of dimensionality. So, we filtered the data prior to feature engineering by *exploratory data analysis (EDA)*, visualizing the distributions of the values, and removing the entries that act as outliers and have very few contributions compared to the distribution of the majority values before going through the encoding process. Also, our target attribute 'twitter_posts_count' is of discrete data type. So, we converted it into 05 classes (class1 to class5) based on the number of shares a twitter post gained. We used equal width binning strategy for this purpose. An example is shown in Table 3.
3. *Model Building:* In this step we built our model using three in-built multi-class classification algorithms: Naive Bayes Classifier, K-nearest neighbor (KNN) algorithm, and Random Forest. We also used one-vs-rest classification for SVM and Xtreme Gradient Boosting (XGBoosting) classifier.

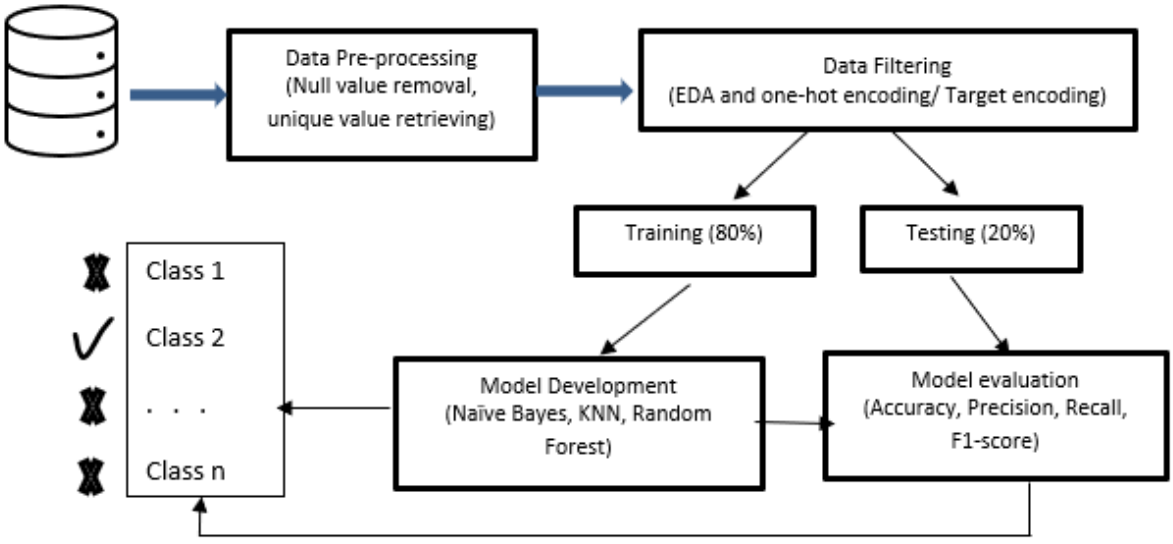

Figure 1: System Architecture

Table 1. Overview of the attributes and preprocessing techniques used

Sl.	Features/Attributes	Original Data Samples	Original Attribute Type	Converted Attribute Type	Preprocessing Technique	Target Class
1	Scopus	['Agricultural and Biological Sciences', 'Life Sciences', 'Immunology and Microbiology']	nominal	binary	Binary encoding (by manual coding)	twitter_posts_count (discrete-> multiclass): equal-width bining (5 classes)
2	twitter_poster_types	{'member_of_the_public': 1, 'practitioner': 4}	nominal	Numerical	By manual coding	
3	subjects	['acquiredimmunodeficiencysyndrome', 'education']	nominal	binary	Binary encoding (by manual coding)	
4	funders	['aruk', 'mrc', 'wellcome', 'dohuk']	nominal	binary	Binary encoding (by manual coding)	
5	publisher	'American Association on Intellectual and Developmental Disabilities 501 3rd Street, NW Suite 200, Washington, D.C. 20001'	nominal	binary	One-hot encoding	
6	journal	'AACN Advanced Critical Care'	nominal	binary	One-hot encoding	
7	research_type	'article'	nominal	binary	One-hot encoding	
8	author_loc	{'ln': -75.4999, 'country': 'US', 'lt': 43.00035}	nominal	binary	One-hot encoding	
9	twitter_author_followers	100 (int type)	discrete	numerical	Log2 transform	
10	tweet_posted_on	'2017-03-29T05:56:19+00:00 '	string	float (only year)	By manual coding	

Table 2. Preprocessing example

Attribute	**Current values**	**After primary preprocessing**
subjects	['acquiredimmunodeficiencysyndrome', 'education']	**subject1**:'acquiredimmunodeficiencysyndrome' **subject2**: 'education'
author_loc	{'ln': -75.4999, 'country': 'US', 'lt': 43.00035}	**Country_code**: 'US'
publisher	公益社団法人 日本薬学会	removed entry (non-english)

Table 3. Converting target to multiclass

Class Attribute (count)	**Class**
1<=twitter_posts_count <=10	Class1
11<=twitter_posts_count <=20	Class2
. . .	. . .
41<=twitter_posts_count <=50	Class 5

Implementation

We implemented the work in windows 10 operating system with 64-bit core-i7 processor. As a programming language we used none other than python as it provides very handful libraries like scikit-learn, pandas, matplotlib, seaborn, etc. for data analysis, visualization, and scientific calculation like building machine learning models

Performance Evaluation

To evaluate our model, we will split our dataset into 80:20 ratio for training: testing. We tested our data to see how accurately it is predicting the classes, i. e., given the attributes of a scholarly article, in which class (range of shares) our model is predicting the article to be which will give us an idea how much attention an article is getting in Twitter. We used four basic machine learning model evaluation techniques: Accuracy, Precision, Recall, and F1-score to evaluate our model and generated classification report for each

model's performance. A comparative analysis of the performance of our dataset with regard to each metric for each of the three classifiers is shown in Table 4. It's obvious from the report that KNN did the best performance even if we didn't used one-vs-rest method for it. As KNN is in-built characterized for multiclass classification, it did the best performance as we expected. Also, decision tree-based ensemble method Random Forest did good job as well in prediction in both with and without one-vs-rest approach. However, SVM showed very poor performance even if we tried to apply it using one-vs-rest classification.

Table 4. Converting target to multiclass

Model used	***Accuracy***	***F1-score***	***Precision***	***Average Recall***
SVM	42.75	25.60	8.6	20
KNN	***81.98***	***81.51***	***77.2***	***72***
Random Forest	73.44	72.11	0.65	58.6
One-vs-Rest (SVM)	42.75	25.60	8.6	20
One-vs-Rest (Xtreme Gradient Boosting)	79.24	78.91	0.71	0.69

Conclusion and Future Work

This project focuses on predicting scholarly articles in twitter. We used the extracted dataset from Twitter and after proper processing could get a standard dataset on which we applied three classification algorithms Naive Bayes Classifier, K-nearest neighbor (KNN) algorithm, and Random Forest. We also tested model performance in one-vs-rest classification strategy for multiclass using SVM and Gradient Boosting algorithm. We found that KNN does better performance than others with a 81.98% accuracy.

In future, we plan to do following amendments in our work:

- Using Neural networks and linear regression.
- Integrating Multiple datasets
- Integrating Twitter text description for prediction, etc.

References:

[1] Haustein, S, Costas, R, Larivière, V. (2015) Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. PLoS ONE 10(3): e0120495.

[2] Smith R. Measuring the Social Impact of Research. BMJ 323(7312): 528 (2001)

[3] Valerio, A., Robin I., Dunbar M. Information Diffusion. In Online Social Networks, 2015

[4] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology, 2009

[5] Boyd, D., Golder, S., Lotan, G.: Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In: 43rd Hawaii International Conf. on System Sciences, pp. 1–10 (2010)

[6] Bandari, R., Asur, S., Huberman, B.: The pulse of news in social media: forecasting popularity. In: ICWSM (2012)

[7] Li, Y., Yu, H., Liu, L.: Predict algorithm of micro-blog retweet scale based on SVM. Application Research of Computers 30(9), 2594–2597 (2013)

[8] Zhang, Y., Lu, R., Yang, Q.: Predicting Retweeting in Microblogs. Journal of Chinese Information Processing 26(4), 109–114 (2012)

[9] Hong L, Dan O, Davison BD. *Predicting Popular Messages in Twitter*. International World Wide Web Conferences. Hyderabad, India. 2011: 57-58.

[10] Jenders M, Kasneci G, Naumann F. *Analyzing and Predicting Viral Tweets*. International World Wide Web Conferences. Rio de Janeiro, Brazil. 2013: 657-664.

[11] Gao S, Ma J, Chen Z. *Modeling and Predicting Retweeting Dynamics on Microblogging Platforms*. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. New York City, USA. 2015:107-116.

[12] Xu Z, Yang Q. *Analyzing User Retweet Behavior on Twitter*. The International Conference on Advances in Social Network Analysis and Mining. Calgary, Canada. 2012: 46-50.

[13] Morchid M, Dufour R, Bousquet P-M, Linares G, Torres-Moreno J-M. Feature Selection using Principal Component Analysis for massive retweet detection. *Pattern Recognition Letters*. 2014; 49: 33-39.

[14] Suh B, Hong L, Pirolli P, Chi EH. *Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network*. 2010 IEEE Second International Conference on Social Computing (SocialCom). Minneapolis, USA. 2010: 177-184.

[15] Chen Z. *Modeling and Predicting Retweeting Dynamics on Microblogging Platforms*. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. New York City, USA. 2015:107-116.

[16] C. Remy, N. Pervin, F. Toriumi, H. Takeda, Information diffusion on twitter: everyone has its chance, but all chances are not equal, in: Signal-Image Technology & Internet-Based Systems (SITIS), 2013 International Conference on, IEEE, 2013, pp. 483–490.

[17] Kwak H, Lee C, Park H, Moon S. *What is Twitter, a Social Network or a News Media?* Proceedings of the 19th International Conference on World Wide Web. Raleigh, North Carolina, USA. 2010: 591-600.

[18] X Zhu, P Turney, D Lemire, A Vellino - Journal of the Association for Information Science and …, 2015

[19] Peng, H., Zhu, J., Piao, D., Yan, R., Zhang, Y.: Retweet Modeling Using Conditional Ran dom Fields. In: ICDM Workshops, pp. 336–343 (2011)

[20] Hansen, D., Shneiderman, B., and Smith, M.A. (2011), Analyzing Social Media Networks with NodeXL: Insights from a Connected World, Morgan Kaufmann, Burlington, MA.

[21] T. B. N. Hoang, J. Mothe “Predicting information diffusion on Twitter – Analysis of predictive features”, Journal of Computational Science, vol. 28, pp. 257-264, September 2018.