DSI DATA SCIENCE AND INFORMETRICS
VOL.2, NO.4, NOV. 2022

Automatically identifying the motivations of Facebook mention to scholarly outputs based on Light GBM

Houqiang Yu[a,b*], Wei Zhang[b], Yue Wang[b], Yinghua Xie[a]

a. School of Information Management, Sun Yat–sen University, Guangzhou, China
b. School of Economics & Management, Nanjing University of Science and Technology, Nanjing, China

ABSTRACT

[Purpose/Significance] The article investigated the automatic identification of the motivation of Facebook mention to scholarly outputs based on Light GBM algorithm, in order to achieve more in-depth usage of Facebook mention on a large scale. [Methodology/Procedure] Based on three types of contextual data, including mentioned scholarly outputs, Facebook users who post scholarly outputs, and text of Facebook posts to scholarly outputs, promising relevant features were extracted, and machine learning algorithms were used to automatically identify the motivations. [Results/Conclusions] (1) Features significantly correlated to the motivation of Facebook mention are identified in all three types of contextual data. In particular, relevant features are the altmetric attention score, the number of collaborative countries, the number of followers, the number of likes, the identities of Facebook users who post scholarly outputs and the number of comments on Facebook posts; (2) The prediction precision of the Light GBM classification model for motivation of Facebook mention was 0.31. In comparison, the classification precision without the text features of Facebook posts was 0.35, which was higher than the overall feature combination. The classification precision with only the post text features was 0.27. After combining the length and language of posts, the precision was improved to 0.30; (3) The classification precision of Facebook motivation has a positive correlation with users' activity. After combining all features, the classification precision of the first quartile users in terms of productivity reached 1, the classification precision of the second quartile was 0.36, and for the third quartile, the classification precision was 0.32. In conclusion, considering the high complexity of automatic classification of motivation of Facebook mentions, the study has achieved relatively high classification precision and could provide reference for future studies.

KEYWORDS

Facebook mention; Facebook mention motivation; Automatic classification; Light GBM

1 Introduction

Due to the diversity of altmetrics data producers, data sources, scholarly dissemination,

* Corresponding Author: yuhq8@sysu.edu.cn

and the way of discussion, the motivation of social media mentions has been more diversified than that of citations (Yu & Li, 2021). The motivation is a major bridge for connecting *data* with *value*, so it is particularly urgent to clarify the motivation of altmetrics data. A substantial array of research has investigated the users' motivation in professional academic online communities till now. In Puschmann and Mahrt's (2012) content analysis study about 339 scholarly discussions, participants from the academia explained that they used social media platforms for online scholarly communication, documentation, promotion, discussion with peers, and education. Kjellberg (2010) divided the purpose of users engaged in academic activities into three categories: information sharing, academic creation, and interaction. Shema et al. (2015) found that the vast majority (90%) of blog users mentioned scholarly output for academic discussion. They might comment on the experimental results, practically apply specific research, or provide practical advice. Some social platforms, such as Twitter and Weibo, have attracted the extensive attention of researchers, but there is little research on Facebook altmetrics. Yu et al. (2017) coded the potential motivation and sentiment distribution of Weibo users, and the results showed that the vast majority (85%) of Weibo posts to scholarly papers were neutral, and more than 90% were for discussion (51%) and dissemination (41%). Na (2015) classified the motivation of Twitter users mentioning psychological outputs and found that more than half (53%) of users were aimed at summarizing research findings, whereas 31% only wanted to retweet the bibliographic information without sentiment or motivation tendency. Na and Years (2017) also conducted a quantitative and Non-numerical analysis to explore the motivation of Facebook users mentioning scholarly outputs in psychology. The results revealed that Facebook users' motivations included discussion and evaluation (20%), practical application (17%), self-promotion (6%), and data source exchange (6%).

The automatic classification of social media mentions to scholarly outputs is critical for the development and application of altmetrics. The automatic classification of citation indicators was relatively mature. Sula and Miller (2014) used the naive Bayes classifier to classify the sentiment of citations. Based on the ensemble learning classifier, Zhang et al. (2019) adopted SVM (linear kernel and radial kernel), decision tree, logistic regression base classification, took different sentences' segmentation granularities into account, filtered the best-cited fragments, and finally realized automatic classification of citations. Ciancarini et al. (2013) introduced the CiTalO tool, which could automatically identify and classify citation content from the formatted documents. Recently, with the development of altmetrics and natural language processing technology, some scholars started to explore the automatic classification of altmetrics data. Based on the topic modeling method, Na and Years (2017) extracted the bibliographic data and Facebook posts of scholarly outputs, and clustered its topics in psychology. However, there was little research on the automatic classification of the motivation of Facebook mention.

In the previous research on the motivation of Facebook mention (Yu et al., 2021), we divided the motivations into five categories based on bottom-up manual coding: *Sharing & Dissemination, Discussion & Evaluation, Promotion & Marketing, Extension & Connecting, and Generalizing & Summarizing.* Based on the machine learning algorithms, this article aims to use Light GBM to automatically classify the motivation of Facebook mention. Due to the small size of the available corpus, it is impossible to construct a training set based on large sample data, and obtain features via feature engineering to build a more scientific classification model. Therefore, this article was an exploratory research. Previous studies have ad-

dressed the motivation of social media mentions to scholarly articles, but are based mainly on content analysis. From these results, we know that motivations for social media mentions are various. However, in order to make use of the findings in a large scale, content analysis is no longer adequate. Our study, to our best knowledge, is the first attempt to automatically identify the motivation by considering different features of the social media mention data.

2 Methodology

2.1 Research Design

In our previous study (Yu et al., 2021), 1879 data records were manually coded using a bottom-up way, which enabled us to identify and label the motivation of Facebook mention. The classification of the motivation of Facebook mentions is based on three types of relevant contextual data: Facebook users who post scholarly outputs, Facebook posts to scholarly outputs, and bibliographic data of mentioned scholarly outputs. In this article, relevant features were extracted from three types of contextual data. The basic rationale is that we try our best to include as much useful information as possible. Drawing lessons from the process of manually judging the motivation in the content analysis, we summarized the relevant information into three categories for each Facebook mention of a scholarly article. They are data about the scholarly article (what is mentioned), data about the Facebook post (how it is mentioned), and data about the author of the Facebook post (it is mentioned by whom).

The process of data analysis is shown in Figure 1. Firstly, the relevant features from the three types of contextual data are listed in Table 1. Secondly, correlation analysis was carried out to find the significant correlation features which have a stronger influence on motivation. The feature set was constructed, and the classified results were obtained by the simulation training of the classifier. Due to the specialty, the text of Facebook posts to scholarly outputs was separately regarded as the test corpus, and the classification results were connected with the above in a matrix to get the final classification results. Finally, the classification results in the test set and the manual coding results were compared. The details are displayed in section 2.3.

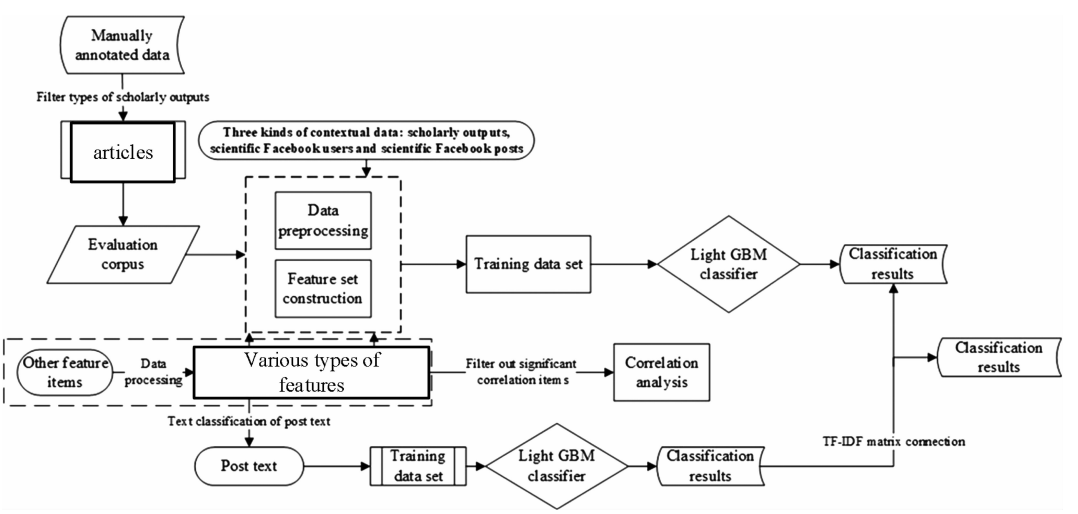

Figure 1 Illustration of research process.

Table 1 Features of Facebook mention contextual data

No.	Feature items	Description
Relevant items of scholarly outputs		
1	Title length	The length of the title of the paper
2	Number of co–authors	The number of co–authors of the paper
3	Number of co–authored institutions	The number of co–authored institutions of the paper
4	Number of co–authored countries	The number of co–authored countries of the paper
5	Subject	The subject of the paper
6	Number of disciplines	The number of disciplines of the paper
7	Number of keywords	The number of keywords of the paper
8	Open access status	Whether the paper is open access
9	Number of pages	The number of pages of the paper
10	Number of downloads	Number of downloads at the retrieval time
11	Citations	Citations of the paper at the retrieval time
12	Number of references	Number of references of the paper
13	Altmetric Attention Score	Altmetric Attention Score of the paper in Altmetric
Relevant data of Facebook users who post scholarly outputs		
14	Type of account	The type of the users´ account is individual/organization
15	User´s identity	User identity types, according to the previous study
16	Number of likes of users	Total number of likes received by the user
17	Number of followers of users	The number of followers of the user
Relevant data of Facebook posts to scholarly outputs		
18	Post content	The full content of the post (including the text in the link preview, the preliminary results of the crawler)
19	Post text	The content of the string in the post
19	Language of post text	The text language of the post (if there are more than two languages, choose the language with more characters as the main language)
20	Number of likes on the post	the number of likes on the post
21	Number of posts reposted	the number of reposts on the post
22	Number of comments	the number of comments on the post
23	Length of post	Length of post text
24	Type of motivation	Type of motivation that is labeled in previous study for each Facebook mention

2.2 Data Collection

The coded data in the motivation research of Facebook mention in this article was collected in Altmetric.com from September 2017 to August 2020, including 29 items, such as titles, output types, altmetric attention scores, DOI, outlets or authors, the types of Facebook users' accounts and identities, and the mentioning motivations. Article is the dominant type of scholarly outputs in the dataset. It has rich bibliographic information and a mature database for access. Therefore, articles (i.e., scholarly papers) are the major research object of this study. Eventually, relevant data from 1605 articles were gathered for analysis.

Bibliographic data of these articles were collected on May 17, 2021, and data of Facebook posts were collected on May 22, 2021.

2.3 Data Processing

After obtaining the raw text of Facebook posts and bibliographic data of articles, in the following procedures, features of three described categories are extracted.

(1) Feature extraction of articles

1. The publications of the labeled data were filtered and extracted to obtain 1605 articles.

2. All articles above were respectively matched with DOI in the Web of Science core collection database, and the full records were exported. However, if the DOI was missing or could not be matched, the article title was retrieved and exported, and the computing code was used to initially extract the article features listed above.

3. A total of 1274 articles were retrieved, while 331 of them were not matched successfully due to missing DOI, long publication history, being conference papers, or online publishing. To solve this problem, Dimensions, Google Scholar, and other databases were used to manually match the features of them. The specific features are shown in Table 1.

(2) Feature extraction of users

1. Python was used to obtain the original mention URL in 1605 publications. Based on the outlet or author, removing duplications was implemented to get the mentioned URL of the deduplicated users' list.

2. The ID in the mentioned URL was extracted by using the Excel column function. Specifically, the users' homepage URL was constructed by adding the URL prefix and the details page suffix to directly access the users' homepage and profile details.

3. For the sake of simulating the manual browsing and waiting for the page to load completely, the selenium was used to control the scrolling of the page slider. And the number of likes and the number of followers of the users was obtained by finding the DIV block of the specific web page format.

4. The loop was set up so that each file contained 15 records of data, to avoid the failure of data collection due to accidental terminal during data crawling. In this way, user's feature data could be obtained.

5. The users' names, types of accounts and identities were matched according to the encoded data in the users' identities research.

(3) Feature extraction of posts

1. The Facebook posts to scholarly outputs' URL was directly obtained from the records obtained from Altmetric database, and the above visiting operation was repeated.

2. The page swiping and the manual browsing were simulated by the selenium module, and waited for the page to fully load.

3. The web page was parsed by invoking the *beautifulsoup* library of *bs4* to obtain the first post mentioning scholarly outputs, the text of posts, the number of likes, comments, sharing text, and reposting.

4. The loop was set, and every 20 records of data were set as a file.

5. The text of posts, the content in the link preview, and the length were extracted or calculated. To identify and correct manually the posts' language, the languid was utilized efficiently to obtain language features.

(4) Correlation analysis

1. The Excel data table, which contained the characteristics of different documents, users,

and posts, was imported into SPSS 25.0.

2. The variable format was checked and revised. In other words, the mistaken scale variables of the Non-numerical features, such as the language and disciplines, were converted into nominal variables.

3. Spearman correlation of all scale variables was eventually calculated, the details can be found in Chapter 3.

(5) Non–numerical data processing

The features of Non-numerical data, such as language, the types of the user's identities, and disciplines, were coded. However, the features were extraordinarily sparse due to the multivariate data. Therefore, the strings were directly replaced with scale variables in this article, then the data pivoting table and VLOOKUP function were used to process and calculate the Spearman correlation of different motivations.

(6) Post text classification

1. The three features, including language, the text, and the length of posts, were loaded. Based on the samples in the users' identities and motivation researches, they were marked as samples 1-5.

2. The language and length of posts in training sets and test sets were fused.

3. Light GBM model was used to classify the text of posts, the number of leaf nodes was set as 50, the learning rate was 0.01, and the number of residual trees was 50.

4. The matrix format of results was exported to convenient subsequent connections.

2.4 Description of Light GBM

Light GBM was an algorithm framework improved by Microsoft Research Asia in 2017 based on the gradient boosting decision tree (GBDT) framework, an ensemble model of decision trees. The GBDT, a popular and widely used machine learning algorithm, trained, iterated, and accumulated multiple weak classifiers by fitting negative gradients to improve their classification performance. However, if the feature dimension was high and the samples were large, the operating efficiency and precision of the GBDT would fail to meet expectations because the traditional boosting algorithm would be time-consuming and inefficient in scanning the data samples of features to locate its optimal cut point.

On the basis of the GBDT algorithm, the GOSS (gradient-based one-side sampling) and the EFB (exclusive feature bundling) technologies were incorporated into Light GBM. For the GOSS, a large part of small gradient data samples was removed, large gradient data samples and a small part of random small gradient data samples were retained to estimate the information gain and the optimal segmentation point, which not only reduced the amount of data but also did not affected the precision. On the other hand, to reduce feature dimensions, the EFP reduced the sparsity of high dimensional features on the basis of binding mutually exclusive features. In short, these two technologies enabled Light GBM model to have more than 20 times faster than the traditional GBDT without reducing the precision (Ke et al., 2017).

Different from the GBDT algorithm sorting with features to find the best segmentation threshold, Light GBM built the histogram decision tree algorithm based on the feature bucketing, that is, continuous float point features were mapped to K discrete items (bin) to form a histogram with K rectangles. In principle, a piecewise function would be performed for each feature in taking eigenvalues, the values of all samples on the feature would be converted into discrete terms, and the continuous values would be also transformed effectively

into discrete values, so that only the height of each rectangle in the histogram, namely the number of bin values, would be calculated statistically. The algorithm successfully transformed #data into #bins, reducing the complexity of the algorithm and improving computational efficiency.

In addition, Light GBM adopted the leaf-wise growth strategy, that is, the nodes with the maximum gain were preferentially selected and divided. Through continuous recursion, resource waste caused by the growth of nodes with a small gain and the training time was reduced, and the work efficiency also was improved without reducing sample weight or learning precision.

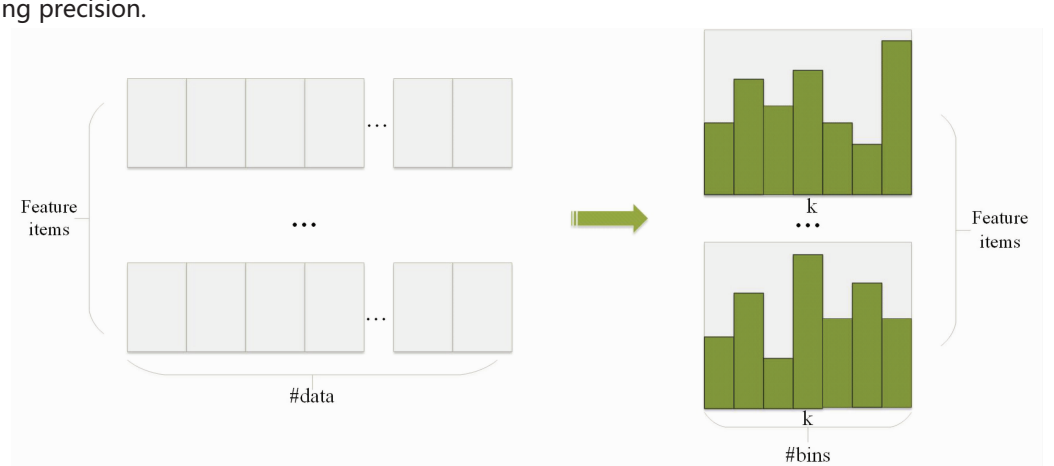

Figure 2 Light GBM histogram algorithm.

The classification process using Light GBM algorithm is shown in Figure 3. The feature data set filtered by correlation analysis and the preprocessed text data were normalized, and the initialization gradient value was calculated to establish the trees. The histogram and the split feature gains were calculated repeatedly within the maximum node range of leaves so that the best split profits were obtained and the root node was established to slice the samples. Based on the above results, the gradient value of the tree was updated until the number of samples was greater than the maximum number of leaves, and leaves could not be divided any further, and all the trees were built. Then the learning rate, the number of residual trees, and other parameters were set to classify and output the above features.

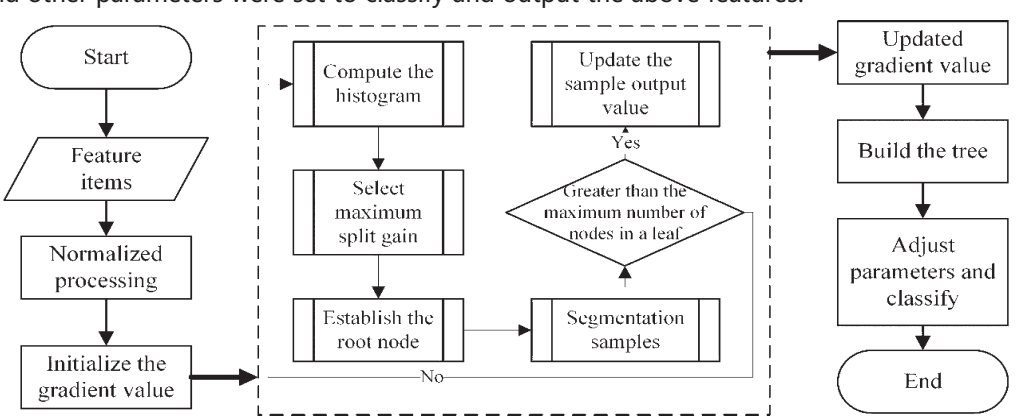

Figure 3 Flow chart of Light GBM algorithm.

3 Classification process and research results

3.1 Basic statistics of feature items

Not every record has features listed in Table 1. For example, not all Facebook accounts had likes, not all publications had been cited or downloaded, and not all Facebook posts were commented, liked, or retweeted. Therefore, in the extraction of feature items, coverage statistics were carried out for all listed features, and the results are shown in Table 2. The coverage rate of most items was relatively high, especially the number of the interdisciplinary, institutions, countries and the other 6 basic indicators was 100%. In other words, each data contained these 9 features. The coverage rate of the types of accounts and users' identities was 93.7%, that is, more than 100 pieces of data could not judge the account and specific identity due to users' privacy settings or incomplete information. The coverage rate of the five feature items, such as the length, the text, and the language of posts, was 89.2%, that is, there were 1432 effective original scientific posts in this study. Among all the features, the lowest coverage rate was distributed in the number of references, keywords, and other 3 features because these features were obtained from the WOS database.

Table 2 Basic statistics of different features

No.	Feature items	Amount	Coverage (%)	No.	Feature items	Amount	Coverage (%)
1	Type of motivation	1605	100.0%	13	The number of institutions in the paper	1605	100.0%
2	Number of users´ followers	1288	80.2%	14	Number of countries in the paper	1605	100.0%
3	Number of users´ likes	1288	80.2%	15	Number of authors of a paper	1605	100.0%
4	Users´ account type	1504	93.7%	16	The title length of the paper	1605	100.0%
5	Altmetric Attention Score	1605	100.0%	17	Length of post	1432	89.2%
6	Number of references	1273	79.3%	18	User identity type	1504	93.7%
7	Paper citations	1273	79.3%	19	The subject of the paper	1605	100.0%
8	Number of downloads	1273	79.3%	20	Number of likes on a post	1430	89.1%
9	Pages of papers	1273	79.3%	21	Number of comments on posts	1432	89.2%
10	Open access status	1605	100.0%	22	Number of reposts	1432	89.2%
11	Number of Keywords	1273	79.3%	23	Language of the post	1432	89.2%
12	Interdisciplinary number of papers	1605	100.0%	24	Post text	1432	89.2%

3.2 Spearman correlation results

As mentioned in chapter 2.3, to improve the convergence of the features, correlation analysis was conducted on the above quantitative features and the processed Non-numerical

features in Table 3. In the features associated with scholarly outputs, the number of collaborative countries and altmetric attention score were significantly correlated with motivation. In the features of Facebook users who post scholarly outputs, the number of followers, the number of likes, the types of users' identities, and the number of comments exerted a great impact on the motivation.

Table 3 Spearman correlation results between feature items and Facebook motivation

Feature items	Correlation	r	Feature items	Correlation	r
Motivation type (tag)	r	1.000	Interdisciplinary number of papers	r	.047
	Sig.(Double tail)	.		Sig.(Double tail)	.058
Number of users´	r	–.060*	The subject of the paper	r	.004
	Sig.(Double tail)	.031		Sig.(Double tail)	.865
Number of users´ likes	r	–.061*	The number of institutions in the paper	r	.036
	Sig.(Double tail)	.030		Sig.(Double tail)	.154
Users´ account type	r	.027	Number of countries in the paper	r	.058*
	Sig.(Double tail)	.295		Sig.(Double tail)	.021
User identity type	r	–.061*	Number of authors of a paper	r	.038
	Sig.(Double tail)	.018		Sig.(Double tail)	.133
Altmetric Attention Score	r	.054*	The title length of the paper	r	.012
	Sig.(Double tail)	.031		Sig.(Double tail)	.625
Number of references	r	.013	Number of likes on a post	r	–.042
	Sig.(Double tail)	.642		Sig.(Double tail)	.113
Paper citations	r	.051	Number of comments on posts	r	.073**
	Sig.(Double tail)	.070		Sig.(Double tail)	.005
Download of papers	r	.020	Number of posts forwarded	r	.033
	Sig.(Double tail)	.469		Sig.(Double tail)	.218
Pages of papers	r	–.011	Post language	r	.014
	Sig.(Double tail)	.698		Sig.(Double tail)	.609
Open access or not	r	–.012	Length of post	r	–.002
	Sig.(Double tail)	.640		Sig.(Double tail)	.941
Open access or not	r	.020	*At level 0.05 (two –tailed), the correlation was significant.		
	Sig.(Double tail)	.487	**At level 0.01 (two –tailed), the correlation was significant.		

3.3 Classification results of Light GBM

The indicators to evaluate the classification results in this article mainly included precision, recall, and the harmonic mean F1-score. These three indicators were calculated by formulas 1 to 3:

$$\text{Precision}=\frac{\text{The amount of Facebook motivational data accurately identified by the system}}{\text{Total amount of Facebook motivations identified by the system}} \quad (1)$$

$$\text{Recall}=\frac{\text{The amount of Facebook motivational data accurately identified by the system}}{\text{Total amount of munually tagged Facebook motivations}} \quad (2)$$

$$F_1\text{-score}=\frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

In the research on the motivation of Facebook mention, the random sampling and stratified sampling strategies were used to divide the source data into five samples. Sample 1 was the overall randomly selected data, including 382 original scientific posts contributed by users. Samples 2-5 were divided in terms of the user's productivity from high to low. Sample 2 consisted of 159 original posts contributed by the top 25% users. The sample 3-5 were divided with 25% users' productivity, including 500, 495, and 343 original posts, respectively. To better compare the classification results of different samples, weighted_avg was used to calculate the test results for different parameters. The weighted average took the proportion of each sample into account, and the precision, recall, and F1-score of each sample were finally averaged. The calculation formulas are as follows:

$$\text{Weighted\_P} = \frac{\sum_{i=1}^{n} (\text{Support\_Precision(i)} / \text{Support\_all})}{N} \quad (4)$$

$$\text{Weighted\_R} = \frac{\sum_{i=1}^{n} (\text{Support\_Recall(i)} / \text{Support\_all})}{N} \quad (5)$$

$$\text{Weighted\_F}_1 = \frac{\sum_{i=1}^{n} (\text{Support\_F}_1\text{(i)} / \text{Support\_all})}{N} \quad (6)$$

(1) Classification results of all features except the text features

Text features refer to features related to text of the Facebook post. For example, the length, the language, the sentiment, the elements (video, figure, emoji, etc.). The number of leaf nodes was set as 50, the learning rate was set as 0.01, and the residual tree was set as 50. The precision, recall, F1-score, and the final weighted_avg of the five samples are shown in Table 4. According to the weighted_avg, the precision was 0.35, the recall rate was 0.35, the F1-score was 0.31, and the overall precision rate was 0.347.

The precision of sample 1 was 0.32, slightly lower than the overall. The precision of the stratified samples was correlated with the activity, and the precision of samples 2-5 decreased successively. The precision of sample 2 was 0.5, and sample 5 was only 0.25, indicating that the higher the user activity, the higher the classification precision. The high productivity users are mainly composed of academic communication or research organizations, so the motivations mostly were sharing or promotion, which could be easily distinguished and classified. With the decrease of users' activity, the motivation became more scattered, and it was hard to judge the discussion evaluation, extended association and summary, thus the machine recognition and classification might face more difficulties.

Table 4 Classification results of different samples (all features except text features)

	Precision	Recall	F1_score
Sample 1	0.32	0.24	0.27
Sample 2	0.50	0.05	0.09
Sample 3	0.37	0.65	0.47
Sample 4	0.33	0.33	0.33
Sample 5	0.25	0.10	0.15
weighted_avg	0.35	0.35	0.31

(2) Classification results of text features

The number of leaf nodes was set as 50, the learning rate was set as 0.03, and the number

of residual trees was set as 50. The precision, recall, F1-score and the final weighted_avg of sample 1-5 are shown in Table 5. According to the weighted_avg, the precision was 0.26, the recall rate was 0.27, the F1 was 0.25, and the overall precision rate of only the post text features was 0.269.

The precision of sample 1 was 0.35, higher than the overall. The precision of sample 2 was 0 because the total number was small, only 159, and most of them were pure links, which both led to the failure to identify significant rules in the text classification model. However, the precision of samples 3-5 still declined, and the correlation between classification precision and users' activity obtained from the above model was still applicable to this classification model.

Table 5 Classification results of text features

	Precision	Recall	F1_score
Sample 1	0.35	0.15	0.21
Sample 2	0.00	0.00	0.00
Sample 3	0.31	0.37	0.33
Sample 4	0.27	0.42	0.33
Sample 5	0.16	0.16	0.16
weighted_avg	0.26	0.27	0.25

(3) The length and language features of the posts were added into the post text classification model

When the parameters remained unchanged, the TF-IDF matrix was used to fuse the length and language of the posts in the feature combination of the training set, and the overall precision was improved from 0.269 to 0.297. The classification results were shown in Table 6. In accordance with the weighted_avg, the precision was 0.27, the recall rate was 0.3, and the F1_score was 0.28. The classification precision of random sampling sample was lower than that in Table 5, but samples 3-5 were improved compared with Table 5, illustrating that the feature combination was superior to the classification model of only the post text features.

Table 6 Classification results of text features adding the length and language features

	Precision	Recall	F1_score
Sample 1	0.22	0.19	0.21
Sample 2	0.00	0.00	0.00
Sample 3	0.36	0.49	0.41
Sample 4	0.27	0.31	0.29
Sample 5	0.27	0.17	0.29
weighted_avg	0.27	0.30	0.28

(4) Classification results of all features

First, the matrix format of the TF-IDF in the training set was exported to facilitate subsequent connection of the features. After trying to fuse all 23 features, the precision was 0.306. According to the weighted_avg, the precision was 0.33, the recall rate was 0.31, and the F1_score was 0.26. Although the precision was lower than the precision without text features,

in Table 7, the classification precision of sample 2 reached 1, sample 5 was 0, and no significant rule was found for sample 5.

Table 7 Classification results of all features.

	Precision	Recall	F1_score
Sample 1	0.21	0.17	0.18
Sample 2	1.00	0.04	0.08
Sample 3	0.36	0.69	0.47
Sample 4	0.32	0.33	0.33
Sample 5	0.00	0.00	0.00
weighted_avg	0.33	0.31	0.26

4 Discussion and Conclusion

4.1 Discussion

To determine the motivation of Facebook mentions, the most intuitive way is to analyze the text of Facebook post. In this regard, text features are perhaps the most important features for automatically identifying the motivation of Facebook mentions. Meanwhile, many other features other than text features can help determine the motivations, including features about users who post scholarly papers on Facebook and the paper that gets mentioned. Therefore, the first model has investigated all features other than text features, and the second model has particularly focused on the text features. Langugage and length of the post, which are external features, do not directly reflect the content of the Facebook post. We have investigated whether these two features would contribute to the identification. The third model has tests on it. In the end, we would like to combine all relevant features to see whether it would bring the best identification results. The fourth model was used for this purpose.

At the beginning of the experimental design, SVM, BP neural network, Adaboost+ decision tree, and Light GBM were all tried to compare the prediction precision. After adjusting parameters, the best classification results of each model were obtained in Table 8. SVM had the poorest prediction effect, and the precision was only 0.26. After reducing the learning rate of BP neural network, the over-fitting problem was minimized, but the best prediction was also only 0.33. Adaboost+ decision tree was selected as the initial weak classifier, and the ensemble learning method was adopted to predict. Finally, the precision was 0.34. The prediction precision of Light GBM was 0.35, which is the best of these models. Besides, in the experiment, the features of Facebook mention had a large dimension and relatively discrete distribution, while Light GBM was more suitable for processing discrete data in high dimensions. Therefore, the model was selected for classification prediction in this study.

Table 8 Feature classification results of different models

	Weighted_Precision	Weighted_Recall	Weighted_F1	Precision
SVM	0.07	0.26	0.11	0.26
BP neural network	0.33	0.04	0.07	0.34
Adaboost+ decision tree	–	–	–	0.34
Light GBM	0.35	0.35	0.31	0.35

In the experiment, correlation analysis was tried to screen the feature items with significant correlation so as to improve the classification precision by feature dimension reduction. However, due to the low correlation, feature dimension reduction could not affect the final low precision, which was 0.27, indicating the problems of data might extremely influence the precision of automatic classification in this study. The best predicted precision of the above experiments was less than 0.4. After verifying the inapplicability of the classification model and the over-fitting of the model, the reason why low precision might be that the distribution of Facebook data was excessively discrete, and it was difficult to find significant rules.

Up to now, research on the automatic classification of motivations was almost blank. Temporarily, the results of this study could not be compared in this vertical field. Based on the relatively mature cited classification, Zhang et al. (2019) implemented the automatic classification study of cited fragments, and the final classification precision was 0.15. Xu et al. (2017) compared the effect of automatic cited fragment recognition on the strength of three methods, in which the word bag model performed best with the precision and recall rates at the lexical level reaching 0.27 and 0.33 respectively, while at the sentence level were only 0.08 and 0.19. In the cited fragment evaluation contest, the best classification result was only 0.15 (Jaidka et al., 2019). At present, the precision of automatic cited fragments recognition was not satisfactory in that the manual annotation and recognition of the cited fragments were arduous, and different understandings between annotators. In addition, the granularity of sentence segmentation and semantic similarity might exert a negative impact on the results, but there were also no unified standards at present. However, the dilemma also existed in the automatic classification of Facebook mention, which required abundant manual work in the initial experimental preparation. For example, in Facebook users' identities and motivations recognition, the subjectivity could not be eradicated. Although the number of sentences and semantic similarity was not a problem in text of posts on Facebook, the features distribution of the contextual data was discrete, so how to balance the richness of indicators and the convergence of data was still a challenge. Therefore, to formulate more reasonable rules, more empirical studies should urgently focus on the automatic classification of the motivation of Facebook mention.

We compared our results with, automatic identification of cited fragments, in fact, to our best knowledge, there is no study on automatic identification for motivation of social media mentions to scholarly papers. Moreover, these studies about automatic identification of cited fragment have also involved with cited scholarly papers and complex cited context, and have adopted similar method for automatic identification. Therefore, they are comparable as regards the method and complexity.

4.2 Conclusion

To fill the current research gap in automatic classification and recognition, this article adopted Light GBM and correlation analysis to automatically classify motivations and discover the particularity of Facebook mention. After turning parameters and optimizations, the following conclusions were drawn:

(1) There were six features in scholarly outputs, Facebook users who post scholarly outputs, and Facebook posts to scholarly outputs significantly related to the motivation, including the altmetrics attention score, the number of collaborative countries, the number of followers, the number of likes, identities of Facebook users who post scholarly outputs, and the number of comments on Facebook posts. The six features were significantly correlated with

the motivation of Facebook mention, that is, the fluctuation in the number and types of these features had a great effect on Facebook motivation.

(2) After integrating all features, the prediction precision of Light GBM for Facebook motivation was 0.31. The precision rate without the text features of Facebook posts was 0.35, which was higher. The classification precision of only the post text features was 0.27. It was improved to 0.30 after combining the length and language of the posts. The reason why the unsatisfactory classification results were that the distribution of Facebook mention might be extremely discrete, and it was difficult to find its own rules.

(3) The classification precision of Facebook motivation was positively associated with users' activity. After integrating all features, the classification precision of sample 2 (i.e., the top 25% of users in productivity) was 1, sample 3 was 0.36, and sample 4 was 0.32. However, the classification precision of sample 5 was 0, indicating that no significant rule was found. This rule was also applicable to the classification model without the text features and the text classification model. This also revealed that Facebook posts to scholarly outputs from highly active users were more regular and more suitable for machine recognition and automatic classification.

There were several limitations in this article. First of all, this study is based on a small sample and is exploratory in nature. Secondly, the improved precision might be scant by changing model algorithms, the feature combination, and the parameters. Finally, the recall rate obtained by the feature combination might also be minor. In future research, it is urgent to further optimize models and rules, improve the classification precision and the recall rate to get more accurate classification results, and promote automatic identification and the large-scale application of Facebook mentions with different motivations.

Acknowledgement

This article is supported by Humanity and Social Science Foundation of Ministry of Education of China (22YJA870016) and National Natural Science Foundation of China (NO. 72274227).

Reference

Ciancarini, P., Iorio, A. D., Nuzzolese, A. G., Peroni, S., & Vitali, F. (2013). Semantic annotation of scholarly documents and citations. *In Congress of the Italian Association for Artificial Intelligence* (pp. 336 –347). Springer, Cham.

Jaidka, K., Yasunaga, M., Chandrasekaran, M. K., Radev, D., & Kan, M. Y. (2019). *The cl–scisumm shared task 2018: Results and key insights.* arXiv preprint arXiv:1909.00764.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS´17)* (pp. 3149– 3157). Curran Associates Inc.

Kjellberg, S. (2010). I am a blogging researcher: Motivations for blogging in a scholarly context. *First Monday, 15* (8). https://doi.org/10.5210/fm.v15i8.2962

Na, J. C. (2015). User motivations for tweeting research articles: A content analysis approach. *In International Conference on Asian Digital Libraries* (pp. 197–208). Springer, Cham.

Na, J. C., & Ye, Y. E. (2017). Content analysis of scholarly discussions of psychological academic articles on Facebook. *Online Information Review, 41* (3), 337–353.

Puschmann, C., & Mahrt, M. (2012). Scholarly blogging: A new form of publishing or science journalism 2.0. *Conference at Düsseldorf*: *Science and the Internet*, 171–181.

Shema, H., Bar–Ilan, J., & Thelwall, M. (2015). How is research blogged? A content analysis approach. *Journal of the Association for Information Science and Technology, 66* (6), 1136–1149.

Sula, C. A., & Miller, M. (2014). Citations, contexts, and humanistic discourse: Toward automatic extraction and classification. *Literary and Linguistic Computing, 29* (3), 452–464.

Xu, J., Li, G., Mao, J., & Ye, G. (2017). Feature analysis and recognition of cited fragments in literature. *Data analysis and knowledge discovery, 1* (11), 37–45.

Yu, H., & Li, L. (2021). Altmetrics mentioned in WeChat articles: Evolution, topic, context and comparison with scholarly publications. *Data Science and Informetrics, 1* (3), 74–92.

Yu, H., Zhang, W., Wang, Y., & Xiao, T. (2021). Who shares scholarly output on Facebook?. In *18th International Conference on Scientometrics and Informetrics* (ISSI 2021) (pp. 1569–1570).

Yu, H., Xu, S., Xiao, T., Hemminger, B. M., & Yang, S. (2017). Global science discussed in local altmetrics: Weibo and its comparison with Twitter. *Journal of informetrics, 11* (2), 466–482.

Zhang, C., Xu, J., & Ma, S. (2019). Research on automatic recognition of cited segments in academic texts. *Information theory and practice, 42* (9), 139.