An Alternative Analysis Method for Measuring the Impact of Academic Papers Shared on Social Media and the Number of Citations Obtained based on a Support Vector Machine Algorithm

by Adian Fatchur Rochim

An Alternative Analysis Method for Measuring the Impact of Academic Papers Shared on Social Media and the Number of Citations Obtained based on a Support Vector Machine Algorithm

Adian Fatchur Rochim[1], Faisal Rizky Rahadian[2] and Dania Eridani[3]

[1]*adian@ce.undip.ac.id*, [2]*faisalrr@students.undip.ac.id*, [3]*dania@ce.undip.ac.id*
Universitas Diponegoro, Faculty of Engineering, Dept. of Computer Engineering, Tembalang 50275,
Semarang (Indonesia)

Abstract

Information technology affects most aspects of human life. Social Media (MedSos) is an information technology product. One of the uses of social media in the academic world is Altmetrics. This indicator is used to measure the impact or influence of social media on indexed papers. Indonesia is one of the countries with the highest social media users in the world. Therefore, this study is proposed to measure the possibility of a correlation between comments / mentions on papers shared on social media and the number of citations obtained. In solving this problem, we propose a method that uses Text Mining to perform Natural Language Processing (NLP) so that machines can understand the meaning of human language and maximize class distance; we used the Support Vector Machine (SVM) algorithm method for classifying opinions in a scientific article. We found that publications shared on social media will have more citations. Papers that have a greater number of positive sentiments will have a large number of citations, whereas the number of tweets on a paper has no effect on the value of positive sentiments and tends to be more contradictory.

Introduction

In its development, information technology is very influential in almost all aspects of life. One example of the results of information technology is social media. Social media is used by people who use it to meet needs, support activities, and open up opportunities to realize new hopes (Akram et al., 2017). The research progress creates this and is one of the developments in communication technology.

Currently, social media has grown rapidly along with technological advances and has penetrated various layers and groups of society. However, there are still few matters concerning social media's impact or influence on researchers' indexed papers in its development. Therefore, sentiment analysis is required regarding this matter based on several factors and criteria. To obtain data with positive or negative review comments that affect the study's h-index or number of citations. Although the h-index has its drawbacks, it is still used as long as there is no better substitute (Rochim et al., 2020).

In previous research, Xiaoli in 2020 discussed dynamics of topic inheritance research and topic innovation by using cross-collection topic models and measuring direct and indirect scientific influence through "citations" (Chen and Han, 2020). Previous research entitled "Sentiment Analysis using a Support Vector Machine" (Nomleni, 2015) discussed the classification of textual documents into several classes, such as positive and negative sentiments and the effects and benefits of sentiment analysis. In this study, the classification of public complaints against the government on social media, Facebook and Twitter, was used with Indonesian language data using a Support Vector Machine (SVM) run in a distributed computer using Hadoop. A study entitled "Adapting SVM for Natural Language Learning: Case Studies Involving Information Extraction" (Li et al., 2008) discussed two techniques to help SVM with the NLP problem's two unique features: unbalanced training data and difficulty obtaining adequate training data. The research problem is how to measure the impact of papers in social media that correlate with several citations. Jason in 2010 stated that Altmetrics (social media in scholars) would become scientometrics 2.0 (Jason et al., 2010). Most of all database indexers, i.e., Scopus, IEEE Xplore used Altmetrics to figure impact profiles of authors from social media.

Methodology

The methodology used in the research adopts an experimental method from various international papers obtained from the Altmetrics website. It discusses how social media's positive comments come to the number of citations in indexed scientific papers made using the Support Vector Machine method to produce some descriptive analysis.
This research begins with data collection totaling 50 different scientific papers from the Altmetrics website (accessed April 13, 2020). This study's data type is primary data as test data obtained from the Altmetrics website and secondary data obtained from Kaggle created by Ali Toosi as training data. The dataset contains 50 datasets of scientific papers, with 700 comments for each dataset. After all the data are collected, the data translation process into English will then be carried out.

Sentiment Analysis Process

Figure 1 illustrates stages of data or documents that enter the system, which are then carried out, cleaning the document to eliminate unnecessary words. After that, the parsing or tokenization process is carried out to divide or break the document into terms based on the stop word and then delete it to filter words or documents. Finally, the stemming process is carried out to obtain common words according to applicable standards.

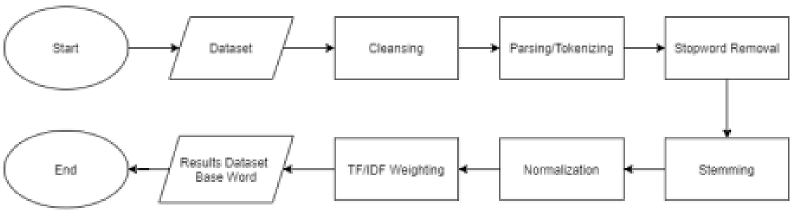

Figure 1. Flow diagram sentiment analysis process

1) Cleansing used aims to remove unnecessary words and characters as noise reduction; Perform cleaning to change all capital letters and documents to lowercase and remove characters other than punctuation, repeating letters, and hyperlinks. Table 1 (Process no.1) presents samples of the cleansing process.

Table 1. Sample of Processing.

No.	Process	Sentence	Results
1.	Cleansing	~~RT @ learn_learning3:~~ It requires large amounts of the image of the person who is the generation source of the video~~.~~	It requires large amounts of the image of the person who is the generation source of the video.
2.	Parsing and Tokenization	but it requires large amounts of the image of the person who is the generation source of the video.	[but, it, requires, large, amounts, of, the, image, of, the, person, who, is, the, generation, source, of, the, video]
3.	Stopword	[~~but, it~~requires, large, amounts, ~~of, the,~~ image, ~~of, the,~~ person, ~~who, is, the,~~ generation, source~~, of, the,~~ video]	[requires, large, amounts, image, person, generation, source, video]
4.	Stemming	[requires, large, **amounts**, image, person, **generation**, source, video]	[require, large, **amount**, image, person, **generate**, source, video]

2) Parsing and tokenization were used to divide large parts of the document into words chop off each word in the text and change all uppercase letters to lowercase. Table 1 (Process no.2) presents samples of the parsing and tokenization process that divides or breaks a large part of a document (sentence) into words; the process chops off each word in the text.

3) Stopword Removal was used to filter words or documents identified as conjunctions, articles, and prepositions. Stopword Removal is the process of removing unnecessary words based on the stopword dictionary in English ('between,' 'yourself,' 'but,' 'again,' 'there,' and so forth). Such words have no meaning. Table 1 (Process no.3) presents the process of stopword removal.

4) Stemming is used to find a root word in a word and remove prefixes, suffixes, and combinations of prefixes and suffixes (Zainuddin et al., 2014). Table 1 (Process no.4) presents samples of stemming that changes the processing of words into basic words by eliminating prefixes, insertions (infixes), suffixes, and combinations of prefixes and suffixes.

5) Normalization is used to measure variable values on a broader scale and change the previous value to become 1. The data should be scaled before training is carried out on the data normalized with mean = 0 and standard deviation = 1. The normalization formula is

$$\text{New Value} = \frac{[Old\ Value(Average)]}{Standard\ Deviation} \quad (1)$$

Data that have been normalized will be divided into training data and testing data using the cross-validation method.

6) Weighting. In weighting this term, the results of the stemming process will be used in calculating the number of documents' Term Frequency (TF), the number of documents that have a term (DF), and the IDF value such as the formula (Nomleni, 2015), (Amrizal et al., 2018). Table 2 presents a sample of the weighting process.

Table 2. Weighting.

Tweet	*Rank*
Ability	0.50
Able	0.69
Work	1.00
Year	0.71

7) TF-IDF Weighting. The method used to find a representation of the value of the data set is trained, and the results form a vector between a document and a word. This method combines a two-weight calculation concept, specifically Term Frequency (TF) and Inverse Document Frequency (IDF). TF determines the word-for-word weight in a document and IDF serves to construct contributions from words into a document. Term Frequency is the frequency of occurrence of words (F) in a sentence (D), and Document Frequency (DF) is the number of sentences in which a word (F) appears (Nomleni, 2015). This word weighting will produce a word weight value, which indicates the importance of each word in the document (Vijayarani et al., 2015). This TF-IDF weighting calculation is formulated in the following equation:

$$\text{idf} = \log\left(\frac{N}{df}\right), \quad (2)$$

where
N = number of document collections, and df = number of documents containing pre-determined term (f)

$$W_{dt} = tfdt \times idft, \quad (3)$$

where

W_{dt} = weight term (t) against the document, (d)

idf = inversed document frequency (log/(N/df)).

$tfdt$ = number of occurrences of term (t) in the document (d), and

Support Vector Machine Algorithm (SVM)

SVM is used as a classification algorithm to find the best hyperplane by maximizing the distance between classes. Classification is done to find a hyperplane or the boundary line (Decision Boundary), which separates a class from another class.

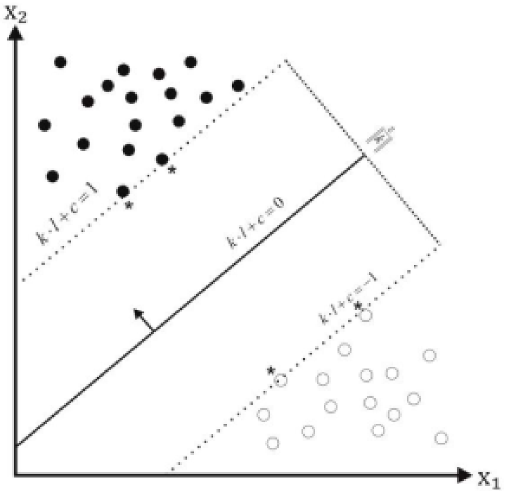

Figure 2. When the wrong hyperplane is positive (+1) and negative (-1).

Figure 2 shows a solid line that is the best hyperplane found by the SVM algorithm, located right in the middle of the second class. The distance between the hyperplane and the data object is different from the near (outermost) class, given an asterisk or star and the data object. The outermost closest to the hyperplane is called the Support Vector (Marafino et al., 2014).

Precision, Recall, and F-Measure

After the SVM classification process is complete, testing of the classification is carried out to measure the performance value of the system that has been created. The formulas used for Precision, Recall, Accuracy, and F-Measure are as follows:

$$\text{Precision} = \frac{TP}{TP+FP}, \quad (4) \qquad \text{Recall} = \frac{TP}{TP+FN}, \quad (5)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \text{and} \quad (6) \qquad \text{F} - \text{Measure} = 2 * \frac{precision*recall}{precision+recall}. \quad (7)$$

Precision is the ratio of true positive predictions compared to total positive predicted results (Irawan et al., 2018), Recall is the ratio of true positive predictions to total true positive data (Flach et al., 2015). Accuracy is the number of correctly predicted documents divided by the total number of documents, and F-Measure is a weighted average comparison of precision and recall (Lipton et al., 2014).

Results and Discussion

From the results of the research that has been conducted, the results obtained from the experiment are Precision, Recall, F-Measure, Accurate, the number of citations, and positive and negative sentiments from each paper.

Table 2 presents the confusion matrix and results of SVM. This method is used to avoid a perfect score but fails to predict anything in the not-yet-visible data (overfitting).

Table 2. Confusion matrix and results of SVM.

Actual Data	*Predict Data*		*Precision*	*Recall*	*F-Measure*	*Accuracy*
	Positive	*Negative*				
Positive	19	6	0.76	1	0.86	0.96
Negative	0	115	0.95	1	0.97	

Positive class:

$$Precision = \frac{TP}{TP + FP} = \frac{19}{19 + 6} = 0.76 \qquad F - Measure = 2 \times \frac{precision \times recall}{precision + recall} = 2 * \frac{0.76 \times 1}{0.76 + 1}$$

$$Recall = \frac{TP}{TP + FN} = \frac{19}{19 + 0} = 1 \qquad = 0.86$$

Negative class:

$$Precision = \frac{TN}{TN+FP} = \frac{115}{115+6} = 0.95$$

$$F - Measure = 2 \times \frac{precision * recall}{precision + recall} = 2 \times \frac{0.95 * 1}{0.95 + 1} = 0.97$$

$$Recall = \frac{TN}{TN + FN} = \frac{115}{115 + 0} = 1$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{19 + 115}{19 + 115 + 6 + 0} = 0.957 \times 100\% = 96\%$$

Test Results

The measurement results for each paper obtained by the average Precision, Recall, and F-Measure using the SVM classification are presented in Table 3.

Table 3. Sentiment analysis results.

Paper id	*Precision*	*Recall*	*F-Measure*	*Accuracy*	*Number of citations*	*P%*	*N%*
1	0.91	0.96	0.93	0.94	881	70.1	29.9
2	0.99	0.94	0.96	0.99	1875	91.6	8.4
3	0.99	0.94	0.96	0.99	354	85.3	14.7
4	0.99	0.87	0.92	0.99	1155	93.8	6.2
5	0.98	0.95	0.96	0.97	3190	78.7	21.3
6	0.99	0.93	0.95	0.99	641	77.1	22.9
7	0.97	0.94	0.95	0.97	799	51.7	48.3
8	0.99	0.93	0.95	0.99	589	90.3	9.3
9	0.97	0.66	0.74	0.95	1502	91.3	8.7
10	0.98	0.83	0.89	0.97	408	86.5	13.5
11	0.96	0.85	0.89	0.94	2721	79.8	20.2
12	0.97	0.80	0.86	0.94	2981	89.1	10.9
13	0.94	0.82	0.86	0.91	3041	76.4	23.6
14	0.93	0.66	0.71	0.87	415	78.4	21.6
15	0.94	0.82	0.86	0.91	701	76.4	23.6
16	0.91	0.75	0.78	0.85	454	76.3	23.7

Paper id	*Precision*	*Recall*	*F-Measure*	*Accuracy*	*Number of citations*	*P%*	*N%*
17	0.90	0.62	0.65	0.82	265	73.9	26.1
18	0.98	0.85	0.90	0.97	4709	90.2	9.8
19	0.98	0.94	0.96	0.98	30	86.2	13.8
20	0.95	0.95	0.95	0.97	1316	78.7	21.3
21	0.97	0.81	0.87	0.94	312	87.2	12.8
22	0.89	0.75	0.77	0.82	485	72.9	27.1
23	0.96	0.86	0.90	0.94	510	79.8	20.2
24	1.00	1.00	1.00	1.00	282	53.1	46.9
25	0.955	0.78	0.83	0.92	403	75.6	26.4
26	0.955	0.725	0.78	0.91	287	80.3	19.7
27	1.00	1.00	1.00	1.00	197	90.9	9.1
28	0.48	0.50	0.49	0.97	516	88.7	11.3
29	0.97	0.85	0.90	0.96	1145	82.5	17.5
30	0.98	0.85	0.90	0.97	312	89.8	10.2
31	0.48	0.50	0.49	0.97	730	95.2	4.8
32	0.98	0.80	0.86	0.97	954	91.4	8.6
33	0.48	0.50	0.49	0.97	151	96.8	3.2
34	0.92	0.66	0.71	0.86	1242	76.8	23.2
35	0.95	0.94	0.95	0.95	510	68.4	31.6
36	0.87	0.86	0.86	0.89	1588	75.0	25.0
37	0.99	0.87	0.92	0.98	906	91.5	8.5
38	0.98	0.81	0.87	0.96	677	88.1	11.9
39	0.96	0.89	0.92	0.94	729	66.6	33.4
40	0.98	0.91	0.94	0.97	21	71.7	28.3
41	0.95	0.87	0.90	0.92	88	64.6	35.4
42	0.97	0.83	0.88	0.95	228	83.1	16.9
43	1.00	1.00	1.00	1.00	235	91.6	8.4
44	0.91	0.89	0.89	0.89	32	63.8	36.2
45	0.89	0.61	0.62	0.79	608	82.8	17.2
46	0.93	0.70	0.75	0.88	757	74.4	25.6
47	0.98	0.75	0.82	0.97	187	89.6	10.4
48	0.90	0.82	0.85	0.93	409	82.4	17.6
49	0.93	0.65	0.69	0.87	1708	88.9	11.1
50	0.97	0.75	0.82	0.96	440	90.2	9.8
Score	0.93	0.82	0.85	0.94	893.52	81.1	18.9

Figure 3 illustrates the comparison of the correlation between the total number of tweets and positive sentiment. In measuring validity between citations data and positive data, we used Spearman's Rank-Order Correlation method. We found that the citations index has a reasonably low correlation. The correlation between citations and positive sentiment is 0.085. Even though this number is very small, it indicates that positive sentiment influences citations. It can be concluded that the higher the positive sentiment, the larger the number of citations. Figure 4 presents a graph between the number of tweets and the sentiment in the opposite direction. Furthermore, using Spearman's Rank-Order Correlation test between the number of tweets and the positive sentiment gave a negative value of −0.183. However, it can be concluded that it does not affect the sentiment score.

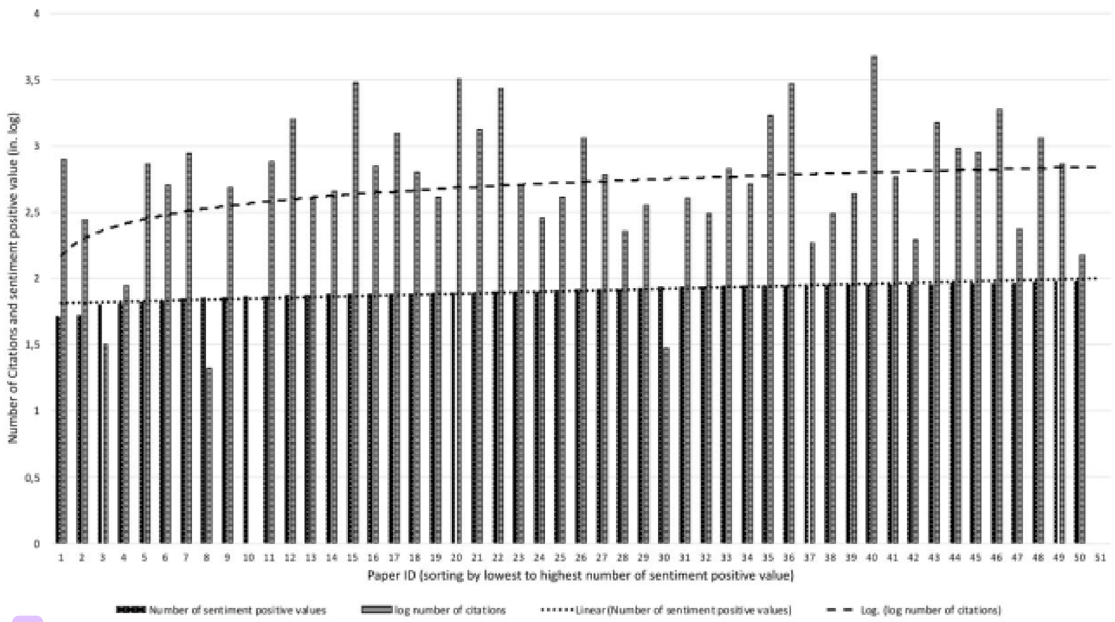

Figure 3. Distribution of the correlation between positive sentiment and number of citations.

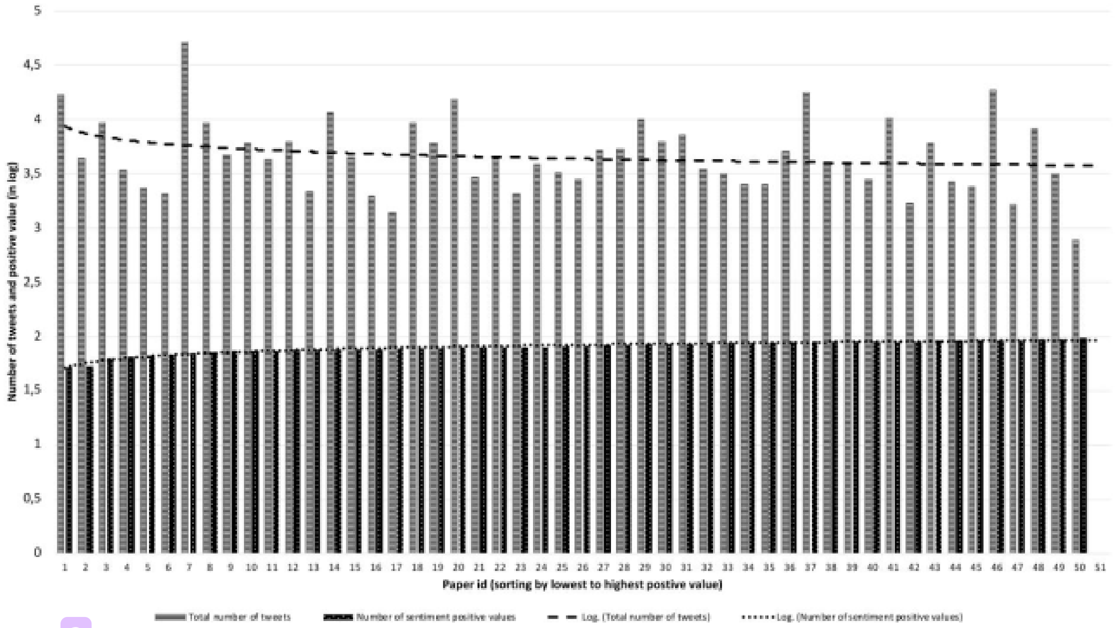

Figure 4. Distribution of the correlation between total tweets and positive sentiment.

Conclusion

We have found that there are correlations among the number of citations of papers obtained, the number of tweets of papers, and number of positive sentiments on papers. We have found that there is a correlation of 0.08 between the number of citations obtained and the number of positive sentiments on a paper. The correlation test between the number of tweets and the number of positive sentiments is −0.183. We can conclude that the number of positive sentiments obtained by the paper will affect the number of citations it gets. The number of tweets obtained by papers on social media has no impact on the number of citations they obtain.

However, this influence is not so significant that it needs to be investigated more. It indicates that the papers that have more several positive sentiments have a larger number of citations.

Acknowledgment

This research was financially supported by the Ministry of Research and Technology, Republic of Indonesia, through Fundamental Research Grant No. B/112/E3/RA.00/2021.

References

Akram, W. & Kumar, R. (2017). A Study on Positive and Negative Effects of Social Media on Society. *International Journal of Computer Sciences and Engineering*, 5 (10), 351–54.

Amrizal, V. (2018). Application of the Term Frequency Inverse Document Frequency (Tf − Idf) and Cosine Similarity Methods in the Information Retrieval System to Determine Web-Based Hadiths (Case Study: Hadith Sahih Bukhari − Muslim*). Journal Teknik Informatika*, 11(2), 149−164, doi:10.15408/jti.v11i2.8623.

Chen, X. & Han, T. (2020). A Micro Perspective of Research Dynamics Through 'Citations of Citations' Topic Analysis. *Journal of Data and Information Science*, 5(4), 1–16.

Flach, P. A. & Kull*,* M. (2015). Precision−Recall−Gain Curves: PR Analysis Done Right. *Neural Information Processing Systems*, 28, 838−846.

Irawan, F. & Samopa. F. (2018). A Comparative Assessment of Random Forest and SVM Algorithms Using Combination of Principal Component Analysis and SMOTE for Accounts A Comparative Assessment of Random Forest and SVM Algorithms, Using Combination of Principal Component Analysis and SM. *The 2nd International Seminar of Contemporary Research on Business & Management 2018*.

Priem, J. & Hemminger. B. (2010). Scientometrics 2.0: New Metrics of Scholarly Impact on the Social Web, *First Monday*, 15 (7).

Li, Y., Bontcheva, K., & Cunningham. H. (2008). Adapting SVM for Natural Language Learning: A Case Study Involving Information Extraction. *Natural Language Engineering*, 15(2),1–25.

Lipton, Z. C., Elkan, C. & Naryanaswamy, B. (2014). Optimal Thresholding of Classifiers to Maximize F1 Measure. *Lecture Notes in Computer Science,* 8725 (2), 225–239, doi:10.1007/978-3-662-44851-9_15.

Manek, A.S., Shenoy, P.D., Mohan, M.C. & Venugolan, K.R. (2017). Aspect Term Extraction for Sentiment Analysis in Large Movie Reviews Using Gini Index Feature Selection Method and SVM Classifier. *World Wide Web*, 20(2), 135–154, doi:10.1007/s11280-015-0381-x.

Marafino, B. J., Davies, J.M., Bardach, M.S., Dean, M.L. & Dudley, R.A. (2014) N-Gram Support Vector Machines for Scalable Procedure and Diagnosis Classification, with Applications to Clinical Free Text Data from the Intensive Care Unit. *Journal of the American Medical Informatics Association*, 21(5), 871–875, doi:10.1136/amiajnl-2014-002694.

Nomleni, P. (2015). Sentiment Analysis using Support Vector Machine. *Thesis, Institute of Technology Sepuluh Nopember*, 2015, pp. 5–6.

Rochim, A. F, Muis, F.A. & Sari, R.F. (2020). A Discrimination Index Based on Jain's Fairness Index to Differentiate Researchers with Identical H-index Values. *Journal of Data and Information Science*, 5(4), 5-18.

Vijayarani, S. & Gurusamy, V. (2015) Preprocessing Techniques for Text Mining: An Overview. *International Journal of Computer Science and Communication Networks*, 5(1), 7–16.

Zainuddin, N. & Selamat, A. (2014). Sentiment Analysis Using Support Vector Machine. *I4CT 2014 - 1st International Conference on Computer, Communications, and Control Technology*, pp. 333–337, doi:10.1109/I4CT.2014.6914200.

An Alternative Analysis Method for Measuring the Impact of Academic Papers Shared on Social Media and the Number of Citations Obtained based on a Support Vector Machine Algorithm

ORIGINALITY REPORT

5%	%	5%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Desdwyatma Wahyu Wibawa, Muhammad Nasrun, Casi Setianingsih. "Sentiment Analysis on User Satisfaction Level of Cellular Data Service Using the K-Nearest Neighbor (K-NN) Algorithm", 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), 2018 Publication	1%
2	"Proceedings of the 2nd International Conference on Electronics, Biomedical Engineering, and Health Informatics", Springer Science and Business Media LLC, 2022 Publication	1%
3	Hongyan Wang, Qiangzi Li, Xin Du, Yewei Lu, Jilei Liu. "Effects of water and heat on growth of winter wheat in the North China Plain", Geocarto International, 2015 Publication	<1%
4	"Advances in Intelligent Networking and Collaborative Systems", Springer Science and Business Media LLC, 2018 Publication	<1%
5	Baichuan Sun, Michael Fernandez, Amanda S. Barnard. "Machine Learning for Silver Nanoparticle Electron Transfer Property Prediction", Journal of Chemical Information and Modeling, 2017	<1%

	Publication	
6	Frank Emmert-Streib, Shailesh Tripathi, Matthias Dehmer. "Analyzing the scholarly literature of digital twin research: Trends, topics and structure", IEEE Access, 2023 Publication	<1 %
7	Yanmeng Li, Huaijiang Sun, Wenzhu Yan, Qiongjie Cui. "R-CTSVM+: Robust capped L1-norm twin support vector machine with privileged information", Information Sciences, 2021 Publication	<1 %
8	Clement Yu. "Discovery of similarity computations of search engines", Proceedings of the ninth international conference on Information and knowledge management - CIKM 00 CIKM 00, 2000 Publication	<1 %
9	IAN RUTHVEN, MOUNIA LALMAS. "A survey on the use of relevance feedback for information access systems", The Knowledge Engineering Review, 2003 Publication	<1 %
10	Fengjie Wang, Mehebub Sahana, Bahareh Pahlevanzadeh, Subodh Chandra Pal et al. "Applying different resampling strategies in machine learning models to predict head-cut gully erosion susceptibility", Alexandria Engineering Journal, 2021 Publication	<1 %
11	Vo Ngoc Phu, Vo Thi Ngoc Chau, Nguyen Duy Dat, Vo Thi Ngoc Tran, Tuan A. Nguyen. "A valences-totaling model for English sentiment classification", Knowledge and Information Systems, 2017 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

An Alternative Analysis Method for Measuring the Impact of Academic Papers Shared on Social Media and the Number of Citations Obtained based on a Support Vector Machine Algorithm

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8