Correlation between Positive Sentiment Twitter Mention and Number of Citations in TOP of 50 Altmetrics

Adian Fatchur Rochim[1], Amara Ranindhita[2], Dania Eridani[3]
[1]adian@ce.undip.ac.id, [2]amarar@student.ce.undip.ac.id, [3]dania@ce.undip.ac.id
Universitas Diponegoro, Faculty of Engineering, Dept. of Computer Engineering, Tembalang 50275, Semarang (Indonesia)

Abstract— **Social media nowadays has become a very popular communication tool. One of the uses of social media in the academic world is Almetrics. Altmetrics (alternative metrics) are used to measure the impact of social media share on indexed papers. Altmetrics calculates the impact of research results on several social media. The number of shares on social media in indexed papers often cannot be assessed for quality and accuracy because it is not through peer review. This study aims to determine the correlation between positive sentiment twitter mention and the number of citations using the Support Vector Machine Algorithm. The result showed a correlation between positive sentiment and the number of citations obtained, which is 0.132, and the correlation between positive sentiment and the number of tweets is -0.0276. Although the result is quite low, it shows that if the positive sentiment is high, the number of citations will increase.**

Keywords—Altmetrics, citation, support vector machine, social media, sentiment analysis.

I. Introduction

Academic papers are papers written to discuss a specific problem based on in-depth research. In academic papers, there is a term called citations, citations are a way of giving credit when certain material in academic papers comes from another source. The number of citations is a parameter of the number of influences obtained by an academic paper. Many things affect the reach of an academic paper; social media is one of them.

Social media have had a significant impact on the academic world, one of the uses of social media in the academic world is Almetrics. Altmetrics (alternative metrics) is a new metric used to calculate the interaction of readers of indexed online papers on social media. Altmetrics considers many factors that can be indicators of the impact and influence of research results.

In 2020, based on We Are Social data, there were 160 million social media users in Indonesia in January 2020.[1]. The increase in discussion forum users can be used as a source of data that can be processed into useful information to find patterns of community behavior towards a particular[2].

However, The number of quotes or shares on social media cannot be assessed for quality and accuracy because they are not through peer review, so further research is needed on this issue.[3]

There are several ways to find out the quality of shares on social media, one of them is by sentiment analysis. Sentiment analysis or opinion mining is an activity that utilizes existing data on social media to find public opinion, as outlined in the text[4]. Sentiment analysis will classify the polarity of the text in a sentence or document to find out whether the opinion expressed in the sentence or document is positive, negative, or neutral[5].

Based on the problems above, this study aims to determine the relationship between the number of positive sentiments on social media with citations obtained by a paper using sentiment analysis. The algorithm used is SVM (Support Vector Machine) to assist or as a reference for other authors in conducting sentiment analysis.

This paper divides into five sections. The first section is an introduction, the second is a literature review, the third is a research methodology, the fourth is results and discussion, and the fifth is the conclusion.

II. Related Work

This research was conducted based on several previous studies with related methods. Rochim, et al. (2020) used the Support Vector Machine algorithm to find that there was a weak correlation between the number of positive sentiments and citations with a value of 0.085, in this study also found no correlation between the number of tweets and the number of positive sentiments with a value of -0.183[6].

Also (2019) Jeong, et al. found that exposure on Twitter had a very large correlation with the citation rate of coloproctology papers. In addition, it is stated that the author of the paper should use social media to accelerate the spread of new ideas so that they are beneficial to the community [7].

In addition, several algorithms are commonly used in the analysis, including the Support Vector Machine. After comparing the accuracy, precision, and recall values, Rahat et al (2019), compared the Support Vector Machine and Naive Bayes algorithms. The result shows that in the case of airline reviews. Support vector machine gave way better results than the Naïve Bayes algorithm [8].

In another study, by Widyaningrum et al. (2021) regarding the comparison of kernels on the Support Vector Machine algorithm, including linear, RBF, and polynomial, it can be concluded that the best kernel on the Support Vector Machine algorithm is a polynomial kernel because it is known that the polynomial kernel gets the highest accuracy value, which is 0.91 compared to the linear kernel 0.86 and 0.90 for the RBF kernel[9].

Reski et al. (2020) examined the best combination of comparisons between training data and test data with data divisions of 70:30, 80:20, and 90:10. the results show that a 90:10 ratio produces the highest accuracy value compared to the others[10].

978-1-6654-5564-0/22/$31.00 ©2022 IEEE

III. Research Methodology

Based on problems regarding the quality of sharing on social media without going through peer review and the low correlation value obtained in previous research, this study aims to find out more about the relationship between the number of shares on social media, the number of positive sentiments and the number of citations obtained by indexed papers. This research methodology section will explain the sequence of research steps, datasets, and methods used for classification.

A. Flowchart

Figure 1 illustrates this research's steps, from data collection to analysis and conclusions.

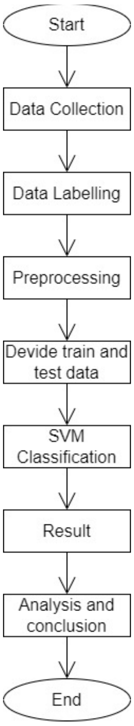

Figure 1. Flowchart steps of this research

B. Data Collection

The first stage is data collection carried out using scraping techniques. The data collected is the tweet from 50 different scientific papers from Almetrics (www.altmetrics.com), with 200 comments for each paper as test data and 1800 data as training data. Data collection of as much as 11,800 data.

The composition of training data and test data in this study is 90:10. Data collection was carried out on June 6, 2022. Each dataset was labeled with positive and negative sentiments. Table 1 is an example of labeled data.

Table 1. Example data

Tweet	Label
You're still the idiot. Here's a study showing animal to human infection from minks. "these persons were infected with strains with an animal sequence signature, providing evidence of animal to human transmission of SARS-CoV-2 w	1
"You're a moron. You act like BBC did study instead of just reporting it. Here's a different direct study showing primates, cats, ferrets, hamsters, rabbits & bats can be infected. Specifically human-animal & animal-human tran"	1
A serious paper one	0
This is a really important study indicating that humans can contract from animals (mink). The fact that humans can infect animals broadens the possibility of viral mutations in animals	0
It's in the original paper. Interesting article on Covid in mink.	0

Manual labeling was performed to classify the dataset into positive sentiment groups and negative sentiment groups. The results of the distribution of labels show that there are 9198 positive sentiments and 2601 negative sentiments as shown in Figure 2.

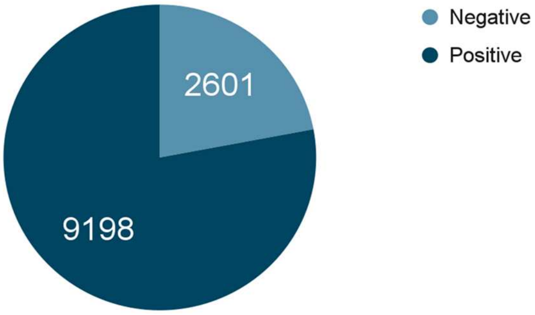

Figure 2. Data source from Altmetrics

C. Preprocessing

Before giving the comparison, preprocessing data will be carried out on the dataset. Preprocessing text aims to prepare unstructured text documents into data that is ready to be used for further processing and to avoid less perfect data, data interruptions, and less consistent data[11]. The stages of text preprocessing are:

- Case Folding is the process of converting all capital letters in a document to lowercase.
- Cleansing is the process of removing numbers, symbols, punctuation, etc from the dataset.
- Normalization is the stage of changing words that are considered nonstandard and can reduce sentiment results, and are replaced with new, more standard words.
- Stopword is the process of eliminating words that often appear but have no special meaning and are

considered unimportant (stopwords) in opinion classification[12].

- Stemming is a text preprocessing process that is used to convert a word in a document into a basic word.
- Tokenization is a process to separate words. These pieces of words are called tokens or terms. The token will be used to support the next process[13].

D. Algorithm Model Design

Support Vector Machine (SVM) is one of the algorithms in machine learning that can analyze data and recognize patterns. This algorithm can be used in the classification process and regression analysis. The basic concept of SVM is to find the best hyperplane that separates the two classes on the input side. Efforts to find the location of this hyperplane are the core of the learning process in Support Vector Machine[14]. Three kernels are most often used in this algorithm, namely the RBF kernel, linear and polynomial. in this study using a polynomial kernel with a comparison of training data and test data 90:10.

E. Model evaluation

The measurement of classification accuracy is carried out to view the report of the classification results on the test data that has been carried out. Confusion matrix is a visualized test formed as a specific table to predict true and false objects. It contains four possible outputs as reference material in comparing the actual events with the predicted events[15]. There are two classes in each report, namely the prediction class and the actual class. There are four forms of reports obtained. First, the value of the number of polarities in the positive class predicted to be the same as the actual class is called the TP (True Positive) value. Two values of the number of polarities in the negative class predicted to be the same as the actual class is called TN (True Negative) values. The three FP (False Positive) values are the number of negative polarities in the actual class that is predicted to have positive polarities. Fourth is the FN (False Negative) value, which is the number of positive polarities in the actual class that is predicted to have negative polarities. Confusion matrix table can be seen in Table 2.

Table 2. Confusion Matrix

		Prediction class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

The confusion matrix table shows reports of predicted and actual data from classification. The classification report will be used to calculate the accuracy. Accuracy compares the total accuracy of the prediction results from positive and negative classes with the total number of predicted classes. Equation (1) is to calculate the values of accuracy.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \quad (1)$$

IV. Result and Discussion

Positive and negative values were obtained from the training data which was used as a classification model, then the test data from 50 papers were classified with the stored model. accuracy values are obtained from formulas (1). The results of positive sentiment, negative sentiment and the accuracy from 50 papers can be seen in Table 3.

Table 3. Sentiment Result

Algoritma SVM Kernel Polynomial							
No	Pos	Neg	Acc	No	Pos	Neg	Acc
1	97.5	2.5	81.5	26	93.0	7.0	85.5
2	95.5	4.5	74.5	27	99.0	1.0	85.0
3	92.5	7.5	81.4	28	98.0	2.0	65.0
4	97.0	3.0	87.4	29	99.0	1.0	81.5
5	98.0	2.0	78.8	30	91.0	9.0	75.0
6	99.0	1.0	82.0	31	99.0	1.0	88.0
7	94.5	5.5	80.0	32	94.5	5.5	65.4
8	96.0	4.0	86.5	33	97.0	3.0	76.5
9	99.0	1.0	77.0	34	99.0	1.0	90.4
10	98.0	2.0	87.0	35	95.0	5.0	75.0
11	98.0	2.0	80.5	36	98.0	2.0	87.4
12	98.5	1.5	73.5	37	96.0	4.0	92.5
13	96.0	4.0	81.5	38	98.5	1.5	93.5
14	95.0	5.0	80.5	39	96.0	4.0	94.4
15	99.5	0.5	77.5	40	96.5	3.5	86.0
16	99.0	1.0	90.0	41	96.5	3.5	92.0
17	99.0	1.0	96.9	42	99.5	0.5	88.0
18	97.0	3.0	66.0	43	97.6	2.4	73.4
19	98.0	2.0	78.0	44	77.3	22.7	94.4
20	99.5	0.5	72.5	45	91.0	9.0	73.3
21	98.0	2.0	88.0	46	99.5	0.5	92.0
22	98.5	1.5	77.5	47	99.5	0.5	92.0
23	87.4	12.6	84.9	48	93.5	6.5	65.5
24	98.5	1.5	84.5	49	98.0	2.0	91.0
25	98.5	1.5	73.0	50	99.1	0.9	91.5

In Table 3 it is known that the highest positive sentiment value is obtained by paper id 15, 46, and 47, the positive sentiment is 99.5 and the highest negative sentiment value is obtained by paper id 44 with a value of 22.7. The average positive sentiment is 96.5 while the negative sentiment is 3.4, based on this data it shows that the top 50 indexed papers used to get a lot of positive sentiment and only a little negative sentiment.

The average accuracy value obtained is 81.45, with the highest accuracy value obtained by paper id 17 of 96.9, and the lowest accuracy value obtained by paper id 28 with a value of 65.0. Based on the accuracy value obtained, it can be seen that the Support Vector Machine algorithm with a polynomial kernel can classify the sentiments in the paper very well, as evidenced by the high curation value obtained.

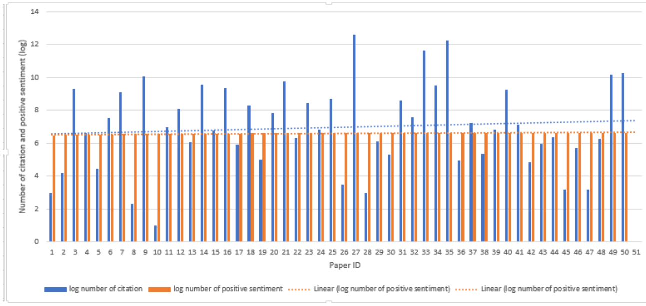

Figure 3. Correlation between positive sentiment and number of citations.

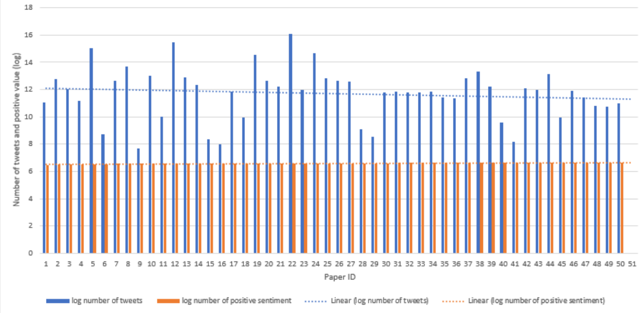

Figure 4. Correlation between positive sentiment and total tweets.

After sorting the paper id based on the sentiment value and calculating the correlation with the number of citations, the results as illustrated by the diagram in Figure 3. This diagram shows the ratio of the number of positive sentiments to the number of citations and it is known that there is a correlation between the number of quotes and the number of positive sentiments is 0.132. Meanwhile, Figure 4 illustrates the comparison between the number of tweets obtained and the number of positive sentiments, and the results show the correlation between the number of tweets and the number of positive sentiments is -0.0276.

V. CONCLUSION

This study concludes that there is a correlation between the number of positive sentiments obtained from a paper on citations, but it does not have a significant effect because it is only 0.132. Although the correlation value obtained is quite low, this proves a correlation between positive sentiment and citations obtained. Based on this value, it is known that the more positive sentiments in a paper, the more citations will be obtained, this proves that using a polynomial kernel can increase the correlation value obtained.

While the number of tweets obtained in a paper does not affect the number of positive sentiments obtained, because the correlation value obtained is -0.0276, this indicates that there is no correlation between the number of tweets and the number of positive sentiments. However, this research also needs to be studied further, and more specifically, regarding other factors that might affect the value of accuracy and correlation.

By clustering the 50 top papers altmetrics, it can be seen that the public has a positive response to the published papers. This statement can be proven by the high average value of positive sentiment obtained, which is 96.5%.

In addition, this study also shows that the high accuracy value is obtained by the Suuport vector machine algorithm with polynomial kernels in classifying existing datasets with an average accuracy value of 81.45%.

For further research, it is recommended to apply neutral sentiment in the classification to obtain a more specific range. Another suggestion is to use more training data and test data to have more variations in the data.

REFERENCES

[1] DATA REPORTAL DIGITAL 2020 : INDONESIA [Online] Available. https://datareportal.com/reports/digital-2020-indonesia

[2] Garg, M., & Kanjilal, U. (2019). A Framework to Process Text Data of Web Discussion Forums: A Study of LIS Links. *DESIDOC Journal of Library & Information Technology*, *39*(6), 315-321.

[3] Zhang, D., & Earp, B. E. (2020). Correlation between social media posts and academic citations of orthopaedic research. *JAAOS Global Research & Reviews*, *4*(9).

[4] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167

[5] Pang, Bo & Lee, Lillian. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval. 2. 1-135. 10.1561/1500000011

[6] Fatchur Rochim, A. 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.

[7] Jeong, J. W., Kim, M. J., Oh, H. K., Jeong, S., Kim, M. H., Cho, J. R., & Kang, S. B. (2019). The impact of social media on citation rates in coloproctology. *Colorectal Disease*, *21*(10), 1175-1182.

[8] Pamungkas, F. S., & Kharisudin, I. (2021, February). Analisis Sentimen dengan SVM, NAIVE BAYES dan KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 pada Media Sosial Twitter. In *PRISMA, Prosiding Seminar Nasional Matematika* (Vol. 4, pp. 628-634).

[9] Rochim, A. F., Widyaningrum, K., & Eridani, D. (2021, December). Performance Comparison of Support Vector Machine Kernel Functions in Classifying COVID-19 Sentiment. In *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)* (pp. 224-228). IEEE

[10] Candra, R. M., & Rozana, A. N. (2020). Klasifikasi Komentar Bullying pada Instagram Menggunakan Metode K-Nearest Neighbor. *IT Journal Research and Development*, *5*(1), 45-52.

[11] Pratama, Y., Tampubolon, A. R., Sianturi, L. D., Manalu, R. D., & Pangaribuan, D. F. (2019). Implementation of sentiment analysis on Twitter using Naïve Bayes algorithm to know the people responses to debate of DKI Jakarta governor election. In *Journal of Physics: Conference Series* (Vol. 1175, No. 1, p. 012102). IOP Publishing.

[12] Langgeni, D. P., Baizal, Z. K. A., & W, Y. F. A. (2010). Indonesian Language News Article Clustering, 2010(semnasIF), 1–10.

[13] Manning, C. D., Raghavan, P., & Schutze, H. (n.d.). Introduction to Information Retrieval.

[14] A. S. Nugroho, A. B. Witarto, and D. Handoko, "Support Vector Machine," in *IlmuKomputer.Com*, 2003.

[15] Subagio, M. M. (2021). *Perbandingan Feature Selection Information Gain dan Chi Square Pada Klasifikasi Film Berdasarkan Sinopsis Menggunakan Metode Naïve Bayes Classifier* (Doctoral dissertation, Universitas Muhammadiyah Malang).