\odot 2005 – ongoing JATIT & LLS

ISSN: 1992-8645

www.jatit.org



AUTOMATIC NEWS BLOG CLASSIFIER USING IMPROVED K-NEAREST NEIGHBOR AND TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY

IRMA YUNITA¹, SENG HANSUN²

^{1,2}Informatics Department, Universitas Multimedia Nusantara, Tangerang, Indonesia E-mail: ¹irma.forworks@gmail.com, ²hansun@umn.ac.id

ABSTRACT

The development of internet technology increases the need for information management for the public. One of the information forms on the internet is the web page. The increment of web pages which is caused by blog content writing has been simplified by the existence of Content Management System, such as WordPress. However, the information is still needed to be organized orderly so that public can easily get relevant information they are looking for. One solution that can be used to organize information is by using text classification. Thus, the objective of this study is to build a WordPress plugin, namely News Blog Classifier that can help users in classifying their articles automatically. Two different algorithms were used to classify the blog content automatically into Health, Economics, Sports, and Technology category, i.e. K-Nearest Neighbor algorithm and TF-IDF. Furthermore, K-Nearest Neighbor algorithm is improvised by adding Cosine Similarity calculation. Based on the test results, the highest precision value is 0.92 obtained from Health category, the highest recall value is 0.97 from Economics category, and the highest F-measure score is 0.88 from Economics category. Overall, from this study, an automatic text classification plugin for news blog content on WordPress CMS has successfully built that can help bloggers in classifying their news articles automatically. In addition to that, the result from the testing phase has discovered a threshold value for each of the categories used in this study that can be used for further research.

Keywords: Content Management System, K-Nearest Neighbor, Text Classification, TF-IDF, WordPress

1. INTRODUCTION

The development of internet technology increases the need of information management in order that public can easily get relevant information [1]. One of the information on the internet is the web page which can be organized by performing text classification.

Text classification is a process or task to automatically group a collection of documents into a category based on a pre-define data set [2]. Text or document classification is one of the main tasks in text mining and information retrieval [3]. By classifying text into certain categories, public can search and find information from internet faster and more efficient [4]. Text classification is an important part of text mining which uses document collection already labeled with certain category [5]. Text classification can be done in three ways, i.e. supervised text classification, unsupervised text classification, and semi-supervised text classification [1]. Supervised text classification uses labeled documents by zero or more categories [1].

Besides, unsupervised text classification takes unlabeled documents to cluster the documents [1]. For semi-supervised text-classification, unlabeled data are used along with few labeled data to classify new unlabeled text documents [1].

Text classification problem can be solved by using K-Nearest Neighbor algorithm (K-NN algorithm) and Term Frequency – Inverse Document Frequency (TF-IDF) [6]. K-NN algorithm is one of the supervised text classification algorithms [1]. Although it is an effective method, K-NN is less efficient classification method due to its high computational procedure [7], [8]. Miah [9] improvised K-NN algorithm by adding Cosine Similarity calculation before calculating the Euclidean Distance. This improvisation is intended to improve application performance to be more efficient because it will only perform further checks on documents that have higher similarity value [9].

One of the popular activities on the web is writing blogs [10], such as news blogging. Detik is an online news portal in Indonesia. Detik ranks fourth overall websites and ranks first as online news

<u>www.jatit.org</u>

4203

E-ISSN: 1817-3195

portal in Indonesia based on visitor traffic according to Alexa [11].

Text classification can make information searching be done easily for electronic documents that are always on the increase. The increment of electronic documents is caused by writing of blog content which is simplified by the existence of Content Management System (CMS), such as WordPress, which is the most widely used CMS [12]. Hence, large text collections can be produced and need to be well organized.

Automatic tools for analyzing large text collections are needed to provide relevant information [13]. Some previous researches have supported this claimed as we can see from the works of Azam et al. [14] who tried to analyze the performance of classification algorithms, i.e. Naïve Bayes and K-Nearest Neighbor, using Scopus dataset; Rivas et al. [15] whose study aimed to propose a novel text classification method on the domain of doctor reviews; Ferdina et al. [16] who had tried to build an automated complaints classification system for student division at Universitas Multimedia Nusantara (UMN) using modified Nazief-Adriani stemming algorithm and Naïve Bayes classifier; and Burel and Alani [17] who had built an open-source web API called CREES that automatically classify social media posts during crisis situations.

Based on the problems describe above, the main research question in this study is how to build a WordPress plugin that implements improved K-NN algorithm and TF-IDF to perform automatic text classification of a blog post in WordPress CMS. The organization of this paper is as follows. Some theoretical basis used in this study will be given in Section 2. It includes the brief explanation of TF-IDF and improved K-NN algorithms incorporated in this study. Some description of WordPress plugin also will be given in this section. In Section 3, an explanation of the research methodology which is divided into two different main processes. i.e. the pre-processing stage and the application plugin phase, is given. The implementation results of TF-IDF and improved K-NN algorithms in automatic news blog classifier as a WordPress plugin will be explained in Section 4, followed by the analysis and discussion in Section 5. Finally, some conclusion will be brought out in the last section of this paper's organization.

2. THEORETICAL BASIS

Two theoretical basis will be explained in this Section, i.e. the improved K-NN and TF-IDF, and WordPress plugin.

2.1 Improved K-NN and TF-IDF

TF-IDF is used to calculate the weight of each word in a document [6]. TF is a calculation of the number of occurrences of words in a document, while IDF is a calculation to reduce the value of TF and increase the weight of a word that rarely appears in a document [6]. Trstenjak et al. improvised TF-IDF calculation by dividing the TF value by the sum of the TF-IDF of a document. The TF-IDF formula is shown on Eq. (1). This improvisation aims to make a better weight value of words in a document that has different word lengths [6].

$$a_{ij} = tf_{ij}idf_i = \frac{f_{ij}}{\sqrt{\sum_{s=1}^{N} \left(tf \ idf(a_{sj})\right)^2}} \times \log_2\left(\frac{N}{df_i}\right)$$
(1)

K-NN algorithm and TF-IDF can perform text classification with high accuracy [6]. However, traditional K-NN algorithm is known to be computationally expensive [18]. Miah [9] improvised K-NN algorithm by adding Cosine Similarity calculation as in Eq. (2).

$$Sim(Q, D_i) = \frac{\sum_i w_{Q,j} w_{i,j}}{\sqrt{\sum_j w_{Q,j}^2} \sqrt{\sum_i w_{i,j}^2}}$$
(2)

K-NN algorithm can perform text classification by using Euclidean Distance calculation [19] as in Eq. (3). Euclidean Distance calculation is done after Cosine Similarity calculation and a category is selected based on the smallest Euclidean Distance value [9].

$$d(x,y) = \sqrt{\sum_{r=1}^{N} (a_{rx} - a_{ry})^2}$$
(3)

2.2 WordPress Plugin

WordPress is the most widely used CMS [12]. WordPress supports the development of client and server side [20]. From the client side, WordPress works well with Hypertext Markup Language (HTML), Cascading Style Sheet (CSS), and JavaScript (JS). WordPress also supports PHP5 and MySQL5 from the server side. Therefore, WordPress can be a good choice for CMS, especially as an editorial site [21]. In fact, it's the most used CMS for blog management [22].

© 2005 – ongoing JATIT & LLS

ISSN: 1992-8645	www.jatit.org	E-ISSN: 1817-3195

The WordPress plugin is used when it comes to developing WordPress functionality with additional features [21]. The process of modifying, setting, and upgrading a WordPress blog can be done easily because of WordPress plugin [23].

3. RESEARCH METHODOLOGY

In this study, pre-processing stage and plugin application are built. Pre-processing stage

aims to process text in news articles from online news portal Detik to create the training word. Plugin application is a plugin for WordPress CMS that is used to classify blog content in text of Bahasa Indonesia. Plugin application uses training word which was generated from the pre-processing stage. The relationship between pre-processing stage and plugin application is shown on Figure 1.



Figure 1: Relationship between pre-processing stage and plugin application

In this study, News Blog Classifier plugin will choose one from four categories, i.e. health, economics, sports, technology. K-NN algorithm has k variable that shows number of neighbors of the document to be classified. The k variable is a simple integer and usually an odd association [9]. For choosing the value of k variable, majority principle is implemented [9]. Since there are four categories that News Blog Classifier can classify, then the value of k variable chosen is five.

Pre-processing stage and plugin application have the same text processing stages: special characters deletion, stopwords removal, and word stemming. The process of removing stopwords uses the list of stopwords in Bahasa Indonesia from the stopwords document by Adikara [24] (hereinafter referred to as 'stopwords document'). Furthermore, the words of the article will be accessed to Kateglo to do stemming words process (using RESTfull API from Kateglo). If we find words that are stopwords, stopwords storage will be updated. In addition, if words are not defined by Kateglo, they will be saved as root words (unique words). Figure 2 shows the stages which have in the pre-processing stage and plugin application.



Figure 2: The text processing stages

3.1 Pre-Processing Stage

Pre-processing stage is a stage used to process text article from Detik (document collections) and creates training word. Training word will be used by plugin application to perform text classification. The number of document collections is 400 documents (100 documents for each category). Figure 3 shows the flowchart of the main process in pre-processing stage.

<u>15th August 2019. Vol.97. No 15</u> © 2005 – ongoing JATIT & LLS

ISSN: 1992-8645

www.jatit.org



E-ISSN: 1817-3195



Figure 3: Flowchart of the main process in pre-processing stage

In Figure 3, the process of exporting stopwords document aims at saving stopwords from stopwords document. Furthermore, document collection is read to extract a document into title and article content. Title and category from document will be saved. Tokenization of article content is used to extract text into word list that will be processed by removing special characters and stopwords, and by stemming words using Kateglo.

3.2 Plugin Application

Plugin application performs text classification by using training word which is created by pre-processing stage. Text classification is performed by choosing one of the four categories (i.e. health, economics, sports, and technology). This plugin is named as News Blog Classifier. News Blog Classifier is used for WordPress CMS. The user interface of this plugin can be seen in the page of creating a new post or editing a post, under the text editor in WordPress CMS. This plugin works if the content post is not empty. Moreover, this plugin allows user to request category, cancel the chosen category, and save post with the selected category.

Improved K-NN Algorithm from Miah [9] and TF-IDF from Trstenjak et al. [6] are implemented in News Blog Classifier. This plugin will work to classify text automatically when user requests category. Figure 4 shows the process when user requests category where improved K-NN Algorithm from Miah and TF-IDF from Trstenjak et al. are implemented.

Journal of Theoretical and Applied Information Technology 15th August 2019. Vol.97. No 15 © 2005 – ongoing JATIT & LLS



E-ISSN: 1817-3195



Figure 4: Flowchart of the main process when user requests category

In Figure 4, text post process is a process to get a list of root words from the article post. Weight for list of root words is done by calculating TF-IDF

from text post. Figure 5 shows the flowchart of TF-IDF calculation from text post.



Figure 5: Flowchart of TF-IDF calculation from text post

15th August 2019. Vol.97. No 15 © 2005 – ongoing JATIT & LLS

ISSN: 1992-8645

www.jatit.org



E-ISSN: 1817-3195

Cosine Similarity calculation is done for all document collections. Flowchart of Cosine Similarity calculation is shown in Figure 6. Five documents with the highest Cosine Similarity value are selected for further Euclidean Distance calculation. Figure 7 shows the flowchart of Euclidean Distance calculation. Choose document process aims at getting a document with the smallest Euclidean Distance. The name of category from the selected document is then displayed to the user.



Figure 6: Flowchart of Cosine Similarity calculation



Figure 7: Flowchart of Euclidean Distance calculation

4. IMPLEMENTATION RESULTS

Training word which is created by preprocessing stage contains 11,068 words that have been accessed to Kateglo, 9,487 root words, 836 stopwords, and number of occurrences of each word on a document. The pre-processing stage results are stored in MySQL and manually exported into a

document with the .sql extension. This document is used by News Blog Classifier plugin to classify text content on WordPress CMS.

The plugin application which was developed is implemented on WordPress 4.7.2. This plugin can be downloaded on https://wordpress.org/plugins/news-blog-classifier/.

<u>15th August 2019. Vol.97. No 15</u> © 2005 – ongoing JATIT & LLS

<u>www.jatit.org</u>	E-ISSN: 1817-3195
2	<u>vww.jatit.org</u>

Once the News Blog Classifier plugin has been downloaded and activated, this plugin is located under the text editor on the page to add a new post and edit a post. The installation can be done by a user who has the privilege as an administrator. Training word will be installed to WordPress system when plugin is installed. Figure 8 shows the main window of the News Blog Classifier plugin.

News Blog Classifier	2. * -
Define Category 0	

Figure 8: The main window of the News Blog Classifier plugin

To classify a blog content automatically, users choose the "Define Category" button. Also, the text editor should not be empty. As an example, Figure 9 shows an article which was placed on the text editor, taken from online Detik news portal in the health category [25].

5 V IN S 8	2012/06 Disc 21
Stroke bukan hanya menyerang orang-orang be	erusia di atas paruh baya. Kaum
muda pun bisa mengalaminya, bahkan pada kor	idisi tertentu stroke juga
ditemukan pada bayi. "Biasanya anak lahir nang	is terus, tahu-tahu lumpuh. Di-
CT scan, ternyata ada perdarahan di otak," kata	i Dr. <mark>d</mark> r Wismaji Sadewo, SpBS
dari RS Premier Bintaro, dalam diskusi di Hotel	The Sultan, Jakarta. Ada
berbagai penyebab serangan stroke pada bayi.	Salah satunya seperti
disampaikan dr Wismaji adalah kurang nutrisi te	ertentu semasa kehamilan,
khususnya vitamin K. Ini terjadi pada kasus APC	CD (Acquired Prothrombin
Complex Deficiency), "Akibatnya pembuluh dara	ah menjadi rapuh, mudah
pecah," katanya. Kasus semacam itu, menurut o	dr Wismaji terjadi pada ibu
hamil yang jarang memeriksakan kandunganny	a secara rutin. Akibatnya anak
lahir dengan kondisi yang sulit dideteksi. "Jumla	ih kasus stroke pada anak
sebenarnya sama seperti pada orang dewasa. H	lanya saja pada anak lebih
'tersembunyi', kata dr Wismaji.	

Figure 9: Text editor with an article

After selecting the "Define Category" button, the News Blog Classifier plugin will define the category of the article entered. Figure 10 shows the progress bar of the content analysis process done using improved K-NN and TF-IDF algorithms. After the blog content analysis process is complete, the plugin application will display the category name (Health means 'Kesehatan' in Bahasa Indonesia) of the article to be posted as shown in Figure 11.

<u>15th August 2019. Vol.97. No 15</u> © 2005 – ongoing JATIT & LLS

ISSN: 1992-8645	www.jatit.org	E-ISSN: 1817-3195
News Blog Classifier		*
Define Category		
	Defining Category Please wait.	

Figure 10: The progress bar shown during blog content analysis process

News Blog Classifier	
Define Category	
Category: Kesehatan	
Remove (Cancell, Category	

Figure 11: The category given by the plugin for input of text from the health article

When a user selects the "Publish" or "Update" button, the article post will automatically be defined with the result of the plugin. The user can choose "Remove (Cancel) Category" to delete a category which was selected. Figure 12 shows the final window interface after an article has been classified automatically using the News Blog Classifier plugin.



Figure 12: News Blog Classifier window after the automatic classification is done

5. ANALYSIS AND DISCUSSION

Testing is performed to measure precision, recall, and F-measure of the category given by the plugin. Dataset for testing is built by random selection of 30 articles from Detik for each category.

Testing phase of plugin application aims at getting the value of precision, recall and F-measure in each category. Precision value can be calculated by comparing the number of texts classified as true with all other texts classified by the application as a category. Recall value can be obtained by comparing the number of texts classified as true with all the other texts tested on a category. F-measure score can be calculated using Eq. (4) [26].

$$F_1 = \frac{2P}{P+R} \tag{4}$$

where P refers to the Precision value that can be obtained from

$$P = \frac{True \ Positive}{True \ Positive + Fals \ Positive} \tag{5}$$

and R refers to the Recall value that can be obtained from

$$R = \frac{True Positive}{True Positive + Fals Negative}$$
(6)

CATECODY	IIE AT TH	ECONOMICS	CDODTC	TECHNOLOCY	TOTAL
CATEGORY	HEALIH	ECONOMICS	SPORTS	TECHNOLOGY	IUIAL
Health	22	0	0	2	24
Economics	1	29	0	6	36
Sports	6	0	27	2	35
Technology	1	1	3	20	25

<u>15th August 2019. Vol.97. No 15</u> © 2005 – ongoing JATIT & LLS

					-	
ISSN: 1992-8645		www.jatit.org			E-ISSN: 181	7-3195
						_
Total	30	30	30	30		

Table 1 shows the results of News Blog Classifier on each category. On Health category, application could classify 22 articles out of 30 and classify two articles from Technology category as Health category. From the testing on Economics category, application could classify 29 articles out of 30 articles, and one article from Health category as Economics category and six articles from Technology category as Economics category. On Sports category, application could classify 27 out of 30 articles, and classify six articles from Health category as Sports category and two articles from Technology category as Sports category. Lastly, on Technology category, application could classify 20 out of 30 articles correctly, and classify one article from Health category, one article from Economics category, and three articles from Sports category as Technology category.

The precision, recall, and F-measure values along with the mean of the three values in each category are shown in Table 2. The News Blog Classifier plugin has precision, recall, and Fmeasure values which are quite high, that is 0.823413 for precision, 0.816667 for recall, and 0.812911 for F-measure. Furthermore, Health category got the highest precision value with 0.916667, and Economics category had the highest recall value with 0.966667. The highest F-measure score was achieved by Economics category with 0.878788. The lowest precision value was obtained bv Sports category with 0.771429, while Technology category had the lowest scores both for recall and F-measure with 0.6666667 and 0.727273 values respectively.

Table 2: Precision, Recall, and F-Measure in each category

CATEGORY	PRECISION	RECALL	F-MEASURE
Health	0.92	0.73	0.82
Economics	0.81	0.97	0.88
Sports	0.78	0.9	0.83
Technology	0.8	0.67	0.73
Mean	0.82	0.82	0.81

Based on the test results that have been done, it was found that the News Blog Classifier plugin can work well enough to classify the text content in Bahasa Indonesia on the WordPress CMS. This is indicated by the value of precision, recall, and F-measure which are high enough. In comparison with Trstenjak et al. [6] research's results where its mean category classification value is 0.8125, improved K-NN with TF-IDF method has slightly better classification results.

From the testing of 30 articles in each category, it yielded the threshold value as shown in Table 3. These threshold values can be used in the future research so that if the tested article obtains a higher Euclidean Distance value than the threshold value of a particular category, the article can be classified into Uncategorized.

CATEGORY	MINIMUM	MAXIMUM	AVERAGE THRESHOLD VALUE
Health	0.16859	0.26384	0.21622
Economics	0.14901	0.25833	0.20367
Sports	0.17355	0.26507	0.21931
Technology	0.14088	0.32540	0.23314

6. CONCLUSION

News Blog Classifier plugin was built successfully by implementing improved K-NN algorithm and TF-IDF to classify the text content automatically on WordPress CMS. The number of document collections is 400 articles obtained from each of the 100 news articles in health, economics, sports, and technology categories. The news articles were taken from the online Detik news portal. Document collection is processed in the preprocessing stage and generates 11,068 words that have been accessed to Kateglo, 9,487 root words, 836 keywords, and the root word number of occurrences in a document. The result of the pre-

<u>15th August 2019. Vol.97. No 15</u> © 2005 – ongoing JATIT & LLS



processing stage is training word that is used by News Blog Classifier plugin to classify text blog content in WordPress CMS. The News Blog Classifier plugin can choose one of the four categories (health, economics, sports, and technology) based on a blog's text content.

Based on the test results, News Blog Classifier plugin has precision value, recall, and Fmeasure in sequence, that are 0.823413, 0.816667, and 0.812911. In addition, the health category received the highest precision value with a value of 0.916667, while the highest recall value is obtained by the economics category with a value of 0.966667. The highest F-measure value is also obtained by the economics category with a value of 0.878788.

Test results of News Blog Classifier plugin also obtain the value of threshold for each category. The threshold for the health category is 0.216215426, economics category 0.203670603, sports category 0.219310036, and the technology category is 0.233139363.

In the future, the threshold values that have been found from the test results can be used to determine the category of the article. Certain categories may be selected if they have Euclidean Distance value that is smaller than the threshold value. If the Euclidean Distance value is greater than the threshold value, the article can be classified as Uncategorized. Furthermore, the addition of document collections can be used to increase the value of precision, recall, and F-measure. Besides that, the addition of categories can be done for further research.

REFERENCES:

- S. Fatima and B. Srinivasu, "Text Document categorization using support vector machine," *Int. Res. J. Eng. Technol.*, vol. 4, no. 2, pp. 141–147, 2017.
- [2] F. Sebasitani, "Kateglo API." [Online]. Available: http://www.programmableweb.com/api/kat eglo.
- [3] A. Jain and J. Mandowara, "Text Classification by Combining Text Classifiers to Improve the Efficiency of Classification," *Int. J. Comput. Appl.*, vol. 6, no. 2, pp. 126–129, 2016.
- [4] A. Faraz, "An Elaboration of Text Categorization and Automatic Text Classification Through Mathematical and Graphical Modelling," *Comput. Sci. Eng. An Int. J.*, vol. 5, no. 2/3, pp. 1–11, 2015.
- [5] A. K. S. Tilve and S. N. Jain, "A Survey on Machine Learning Techniques for Text

Classification," Int. J. Eng. Sci. Res. Technol., vol. 6, no. 2, pp. 513–520, 2017.

- [6] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF Based Framework for Text Categorization," *Procedia Eng.*, vol. 69, pp. 1356–1364, 2014.
- [7] M. Huang, R. Lin, S. Huang, and T. Xing, "A Novel Approach for Precipitation Forecast via Improved K-Nearest Neighbor Algorithm," *Adv. Eng. Informatics*, vol. 33, pp. 89–95, 2017.
- [8] S. Jiang, G. Pang, M. Wu, and L. Kuang, "An Improved K-Nearest-Neighbor Algorithm for Text Categorization," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 1503–1509, 2012.
- [9] M. Miah, "Improved k-NN Algorithm for Text Classification," in *Proceedings of the* 2009 International Conference on Data Mining, 2009, pp. 1–7.
- [10] J. T. Pollock, *Semantic Web for Dummies*. Indianapolis: Wiley Publishing, Inc., 2009.
- [11] Alexa, "Top Sites in Indonesia." [Online]. Available: http://www.alexa.com/topsites/countries/ID
- [12] BuiltWith, "CMS Usage Statistics." [Online]. Available: https://trends.builtwith.com/cms.
- [13] A. J. Mohammed, Y. Yusof, and H. Husni, "GF-CLUST: A Nature-Inspired Algorithm for Automatic Text Clustering," *J. Inf. Commun. Technol.*, vol. 15, no. 1, pp. 57– 81, 2016.
- [14] M. Azam, T. Ahmed, F. Sabah, and M. I. Hussain, "Feature Extraction based Text Classification using K-Nearest Neighbor Algorithm," Int. J. Comput. Sci. Netw. Secur., vol. 18, no. 12, pp. 95–101, 2018.
- [15] R. Rivas, N. Montazeri, N. X. Le, and V. Hristidis, "Automatic Classification of Online Doctor Reviews: Evaluation of Text Classifier Algorithms," J. Med. Internet Res., vol. 20, no. 11, pp. 1–15, 2018.
- [16] V. Ferdina, M. B. Kristanda, and S. Hansun, "Automated Complaints Classification using Modified Nazief-Adriani Stemming Algorithm and Naive Bayes Classifier," J. Theor. Appl. Inf. Technol., vol. 97, no. 5, pp. 1604–1614, 2019.
- [17] G. Burel and H. Alani, "Crisis Event Extraction Service (CREES) - Automatic Detection and Classification of Crisisrelated Content on Social Media," in 15th International Conference on Information

© 2005 – ongoing JATIT & LLS



www.jatit.org



E-ISSN: 1817-3195

Systems for Crisis Response and Management, 2018.

- [18] Y. Wu, K. Ianakiev, and V. Govindaraju, "Improved K-Nearest Neighbor Classification," *Pattern Recognit.*, vol. 35, no. 10, pp. 2311–2318, 2002.
- [19] M.-Y. Su, "Using Clustering to Improve the KNN-based Classifiers for Online Anomaly Network Traffic Identification," J. Netw. Comput. Appl., vol. 34, no. 2, pp. 722–730, 2011.
- [20] M. Olinik and R. J. Armitage, *The Wordpress Anthology*. Collingwood: SitePoint Pty. Ltd., 2011.
- [21] T. D. Hedengren, Smashing WordPress: Beyond The Blog. United Kingdom: John Wiley & Sons, Ltd, 2010.
- [22] W. Aprilius, S. Hansun, and D. Gunawan, "Entity annotation WordPress plugin using TAGME technology," *Telkomnika* (*Telecommunication Comput. Electron. Control.*, vol. 15, no. 1, 2017.
- [23] WordPress.org, "Writing a Plugin." [Online]. Available: https://codex.wordpress.org/Writing_a_Plu gin.
- [24] P. P. Adikara, "Kamus Kata Dasar dan Stopword List Bahasa Indonesia," 2012. [Online]. Available: http://www.hikaruyuuki.lecture.ub.ac.id/ka mus-kata-dasar-dan-stopword-list-bahasaindonesia.
- [25] Detik.com, "Kurang Asupan Vitamin K Bisa Picu Stroke pada Bayi," 2017. [Online]. Available: https://health.detik.com/read/2017/03/16/17 0457/3448834/1300/kurang-asupanvitamin-k-bisa-picu-stroke-padabayi?1992205755.
- [26] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. United Kingdom: Cambridge University Press, 2008.