OPEN ACCESS
Indo-JC
Indonesia Journal of Computing
ISSN 2460-9056
socj.telkomuniversity.ac.id/indojc

Ind. Journal on Computing
Vol. 4, Issue. 2, Sept 2019. pp. 199-204
doi:10.21108/indojc.2019.4.2.346

Topic Classification of Islamic Question and Answer Using Naive Bayes Classifier

Naufal Furqan Hardifa [1], Kemas M. Lhaksmana [2], Jondri

School of Computing, Telkom University
Bandung, Indonesia

[1] naufalfurqan@students.telkomuniversity.ac.id

[2] kemasmuslim@telkomuniversity.ac.id

[3] jondri@telkomuniversity.ac.id

Abstract

Topic classification is one of the most important components in an automatic Islamic question-answering system, which is capable of automatically providing the most relevant answers given a question about the Islamic issue. In our research, the Islamic question-answering system to be built collects existing Islamic questions and answers from trusted online Islamic consultation websites. To speed up the search for finding the appropriate answers, each Q & A entry should be classified into a topic. However, the question-answering system cannot directly adopt the topic classes provided by the online Islamic consultation websites, because different websites use different classifications. Since the number of Q & A entries could reach tenth thousands, an automatic topic classification method is required. In this paper, a naive Bayes classifier is implemented to classify Q & A entries. The classifier gives a satisfying result with 0.88 precision, 0.85 recall, 0.86 f1- score and 0.97 accuracy.

Keywords: classification, naive Bayes classifier, Q & A.

Abstrak

Topic classification adalah salah satu komponen paling penting dalam sistem Tanya-Jawab Islam otomatis, yang mampu secara otomatis memberikan jawaban yang paling relevan dari pertanyaan tentang masalah Islam yang diberikan. Dalam penelitian kami, sistem tanya jawab Islam yang akan dibangun mengumpulkan berbagai pertanyaan dan jawaban Islami yang ada dari situs web konsultasi Islam online yang tepercaya. Untuk mempercepat pencarian untuk menemukan jawaban yang sesuai, setiap entri tanya jawab harus diklasifikasikan ke dalam suatu topik. Namun, sistem menjawab pertanyaan tidak dapat secara langsung mengadopsi kelas topik yang disediakan oleh situs web konsultasi Islam online, karena situs web yang berbeda menggunakan klasifikasi yang berbeda. Karena jumlah entri Tanya-Jawab dapat mencapai ribuan, metode klasifikasi topik otomatis diperlukan. Dalam tulisan ini, naive Bayes classifier dipakai untuk mengklasifikasikan entri Tanya-Jawab. Klasifikasi memberikan hasil yang memuaskan dengan precision adalah 0,88.

Kata Kunci: classification, naive Bayes classifier, Q & A.

I. Introduction

ON this research, a model is used to classify question and answers articles on Islamic religion consultation website into several certain topics. The classification system on each website has a different type of classification system, so to help to classify the articles, will be built an automatic system for classifying all the question and answers article topics. In this research only use one example of a website, but for the future implementation will include many other websites. In this research, the classification system uses the naive Bayes classifier. Naive Bayes classifier is used because it is considered

Received on August 2019. Accepted on Sept 2019

quite easy to understand and has been widely used in other related studies. This research is part of long-term research in which a question answering system will be built. Question answering system is a system that allows a user to ask some question using sentences in the form of natural language questions and returns the answer by analyzing and learning the question given [1][2]. However, on this research, the classification system will be built first. This topic is chosen because so many people access internet nowadays to search for something about Islamic religious knowledge, so a program that can classify the topics from question and answers articles will be built as the requirement to build a question answering system which can make it easier to get the articles people look for.

II. Related Work

There are several previous studies which use a similar tech nique with this research. In the research which has done by Fiarni, C., Maharani, H., Pratama, R. [7], had been developed a sentiment analysis system that allows processing the opinion from social media using text mining. This research uses naive Bayes classifier and its ability to classify the user opinion with the accuracy of 89.21%. In other research which has done by Chandrasekar, P., Qian, K. [8], said that the naive Bayes classifier is simple and effective for classification and it is suggested the most effective method. In the next other research as well, which has done by Qiong Wang, George M. Garrity, James M. Tiedje, and James R. Cole [9] about classifica tion of bacterial 16S rRNA sequences into the new higher- order taxonomy proposed in Bergeys Taxonomy Outline of the Prokaryotes (2nd ed., release 5.0, Springer-Verlag, New York, NY, 2004). In this research, the classification is done using naive Bayes classifier as well where provides taxonomic assignments from domain to genus with the accuracy of 98%.

Naive Bayes is one part of machine learning where it is one part of supervised learning. Naive Bayes determines the probabilities with certain term and conditions. The following bellow is general forms of naive Bayes theorem:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{1}$$

While the naive Bayes classifier is a kind of classifier that uses multinomial distribution [4]. Naive Bayes classifier predicts the probabilities of membership for each class and it has strong independence assumption [6]. Naive Bayes classifier specifies the most likely class to the given example described by the feature vector [5]. In this classification itself, can assume feature which is an independent class, that is $P(\mathbf{X}|C) = \Pi_{i=1}^{n} P(X_i|C)$, where $\mathbf{X} = (X_1, ..., X_n)$ is feature vector.

Before that, in the classification process, the data or file that will be processed into training first. In this process, will be built vocabulary from the data given. From the vocabulary which has been built, words will be chosen with the frequency greater than the cutoff and will be decided to be the feature which will be feature vector or words vector later. After finishing the training process, the next is the classification process.

In the classification process, what is done is to find the maximum value from [10]:

$$\hat{c} = \arg\max_{c \in C} P(c|f_1, f_2, ..., f_n) \tag{2}$$

With applying the Bayes theorem, (2) can be written by [10]:

$$\hat{c} = \arg\max_{c \in C} \frac{P(f_1, f_2, ..., f_n|c)\, P(c)}{P(f_1, f_2, ..., f_n)} \tag{3}$$

Since the value of $P(c|f_1, f_2, ..., f_n)$ for all c have the same size then its value can be ignored, so (3) will be [10]:

$$\hat{c} = \arg\max_{c \in C} P(f_1, f_2, ..., f_n|c)\, P(c) \tag{4}$$

By assuming that every word in $< f_1, f_2, ..., f_n >$ is independent, then $P(f_1, f_2, ..., f_n|c)$ in (4) is written in [10]:

$$P(f_1, f_2, ..., f_n|c) = \prod_i P(f_i|c) \tag{5}$$

And with the final state, the equation can be written by [10]:

$$c_{NB} = \underset{c \in C}{\arg\max}\, P(c) \prod_{i} P(f_i|c) \tag{6}$$

The $P(c)$ value is determined on training, which the value is approached by [10]:

$$P(c) = \frac{N_c}{N_{doc}} \tag{7}$$

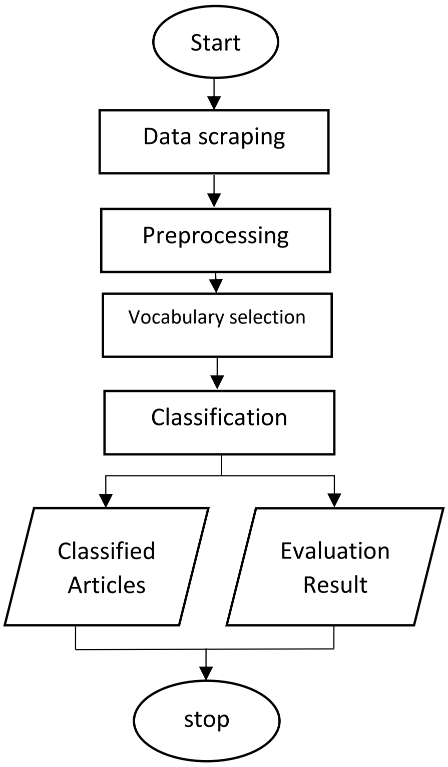
where the $|N_c|$ is the number of documents that have c category in training, while $|N_{doc}|$ is the number of documents in the example used for training.

In this program on this research, the accuracy is counted by using the confusion matrix:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

III. Research Method

Figure 1 is the flowchart of naive Bayes classifier program.



Gambar 1. Program Flowchart.

A. Data Scraping

Collecting data is done by scrapping the data from its websites directly. Scraping or web scraping is a technique or method which is used to gain the information automatically, the purpose is to gain the information and then extract it [3]. The website which is used as references in this research is www.rumahfiqih.com. The data used is in the form of consultation articles about Islam religion. The amount of the data used is 800 question and answer articles, where there are eight categories with each category consist of 100 articles. So, there are 800 links of the website that has been taken for scraping data on this research.

B. Preprocessing

At this step, raw data are processed in the preprocessing process. This process allows raw data to be processed faster in the program because the data has already been simplified. Before the data is processed further, data will be preprocessed first so it can be processed faster in the main program. There is some example of the preprocessing process which has been done, such as stemming, case folding, tokenization, etc. Some preprocessing processes which are used in this research include:

- Stemming is a function for converting the words in a sentence into its basics form.
- Case folding is a fuction for converting the uppercase into lowercase.
- Tokenization is a function for cutting of a sequence of characters into pieces of words or characters.
- Repeated character normalization is a function for removing the repeated characters.
- Enter normalization is a function for turning the text with several lines into one line.
- URL normalization is a function for removing a link from the text.
- Emoticon normalization is a function for normalizing the characters that form emoticons.
- Repeated dot is a function for removing the excessive dots from the text.
- Ellipsis normalization is a function for removing the ellipsis sign from text.

C. Vocabulary Selection

The next step is done building vocabulary based on the data that will be used to be processed. On this step, the words will be filtered and taken the words that are needed only.

D. Classification

After going through the training process before, in this process, the data begin to be classified using a model that has been determined. In this research case, a model that is used is naive Bayes classifier.

E. Evaluation Result

After the data has been classified, the result will be in the form of precision, recall, and f1-score. The precision is a fraction of documents retrieved in a search that is relevant to a query [11] or how much the accuracy of the information requested by the information given. The recall is measuring the percentage of total matching documents in a collection that are returned [11] or how much the success of the model to look for information. While the f1-score is to find out the balance of the value between precision and recall. Then, will be determined the accuracy to count how accurate the result given by the model.

IV. Results and Discussion

A. Test Results

The following table I is the confusion matrix result of the classification.

The following tables II is the confusion matrix result of the classification for each class.

Based on the classification result on table III using the naive Bayes classifier, precision, recall, f1-score were obtained. Where the score of precision is 0.88, recall is 0.85, and f1-score is 0.86.

From the accuracy result on table IV, the classification result get a very good score of accuracy. From the table we get score of 0.97 of the total accuracy.

B. Analysis of The Test Results

From the test results is obtained some measurement results such as precision, recall, f1-score, and accuracy. Where the precision, recall, and f1-score have a score of 0.88, 0.85, 0.86 in a sequence where

Tabel I
Confusion Matrix

Aqeedah	**23**	3	0	0	0	0	1	0
Contemporary	2	**12**	0	1	0	0	1	0
Heritage	3	1	**25**	0	0	0	0	0
Muamalah	0	2	0	**20**	0	0	0	2
Marriage	0	1	0	1	**31**	0	0	0
Fasting	0	0	0	0	0	**17**	1	1
Prayer	1	1	0	0	0	0	**22**	0
Ablution	0	4	0	1	0	1	1	**21**
	Aqeedah	Contemporary	Heritage	Muamalah	Marriage	Fasting	Prayer	Ablution

Tabel II
Contingency Tables

[Aqeedah]	23	4	[Contemp.]	12	4	[Heritage]	25	4	[Muamal.]	20	4
	6	167		12	172		0	171		3	173
[Marriage]	31	2	[Fasting]	17	2	[Prayer]	22	2	[Ablution]	21	7
	0	167		1	180		4	172		3	169

Tabel III
Classification Results

Category	***Precision***	***Recall***	***f1-score***
Aqeedah	0.79	0.85	0.82
Contemporary	0.50	0.75	0.60
Heritage	1.00	0.86	0.93
Muamalah	0.78	0.83	0.85
Marriage	1.00	0.94	0.97
Fasting	0.94	0.89	0.92
Prayer	0.85	0.92	0.88
Ablution	0.88	0.75	0.81
Avg/total	***0.88***	***0.85***	***0.86***

Tabel IV
Results of Accuracy

Category	***Accuracy***
Aqeedah	0.95
Contemporary	0.92
Heritage	0.98
Muamalah	0.97
Marriage	0.99
Fasting	0.99
Prayer	0.97
Ablution	0.95
Avg/total	***0.97***

from the three values can be known that the model can give the requested information quite well and quite well in the accuracy of finding information as well. The average/total accuracy obtained from this model is 0.97. From the accuracy can be known that the prediction score with the actual score has a high level of similarity. Since the data used is also limited where using only one website, so the model has already given quite well results.

V. Conclusion

In this research, we have done classification of question and answers articles about Islam religion issues using naive Bayes classifier. Where the results obtained using this naive Bayes classifier are the precision score of 0.88, recall score of 0.85, f1- score of 0.86 and the accuracy score of 0.97. Therefore, it can be concluded that the naive Bayes classifier can be used for classification.

However, in this research still using one example of a website. Therefore, to obtain higher accuracy, in the future research will be used more websites, so the classification result will gain more optimal value.

Pustaka

[1] Wei, Z., Xuan, Z., Junjie, C. (2012). Design and implementation of influenza Question Answering System based on multi-strategies. 2012 IEEE International Conference on Computer Science and Automation Engineering (CSAE).

[2] Samadi, A., Hanaa, E. F., Qbadou, M., Youssfi, M., Akef, F. (2018). A syntactic and semantic multi-agent based question answering system for collaborative e-learning. 2018 4th International Conference on Optimization and Applications (ICOA).

[3] Vargiu, Eloisa. Urru, Mirko. 2013. Exploiting Web Scraping in a Collaborative Filtering Based Approach to Web Advertising. Italy. Barcelona Digital Technology Center.

[4] Pane, R. A., Mubarok, M. S., Huda, N. S., Adiwijaya. (2018). A Multi-Lable Classification on Topics of Quranic Verses in English Translation Using Multinomial Naive Bayes. 2018 6th International Conference on Information and Communication Technology (ICoICT).

[5] Rish, I. 2001. An empirical study of the naive Bayes classifier. United States: T.J. Watson Research Center.

[6] Mori, T., Tamura, S., Kakui, S. (2013). Incremental Estimation of Project Failure Risk with Naive Bayes Classifier. 2013 ACM / IEEE International Symposium on Empirical Software Engineering and Measurement.

[7] Fiarni, C., Maharani, H., Pratama, R. (2016). Sentiment analysis system for Indonesia online retail shop review using hierarchy Naive Bayes technique. 2016 4th International Conference on Information and Communication Technology (ICoICT).

[8] Chandrasekar, P., Qian, K. (2016). The Impact of Data Preprocessing on the Performance of a Naive Bayes Classifier. 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC).

[9] Wang, Qiong. Garrity, George M. Tiedje, James M. Cole, James R. 2007. Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy.

[10] Jurafsky, Daniel. H Martin, James. 2016. Naive Bayes and Sentiment Classification.

[11] Feldman, S. E. (2012). The Answer Machine. Synthesis Lectures on Information Concepts, Retrieval, and Services, 4(3), 1−137.