Classification of hadith into positive suggestion, negative suggestion, and information

IOP Conf. Series: Journal of Physics: Conf. Series **971** (2018) 012046 doi:10.1088/1742-6596/971/1/012046

Classification of hadith into positive suggestion, negative suggestion, and information

Said Al Faraby, Eliza Riviera Rachmawati Jasin, Andina Kusumaningrum, Adiwijaya

School of Computing, Telkom University

E-mail: saidalfaraby@telkomuniversity.ac.id, eliza@jasin.net, andinakusumaningrum@gmail.com, adiwijaya@telkomuniversity.ac.id

Abstract. As one of the Muslim life guidelines, based on the meaning of its sentence(s), a hadith can be viewed as a suggestion for doing something, or a suggestion for not doing something, or just information without any suggestion. In this paper, we tried to classify the Bahasa translation of hadith into the three categories using machine learning approach. We tried stemming and stopword removal in preprocessing, and TF-IDF of unigram, bigram, and trigram as the extracted features. As the classifier, we compared between SVM and Neural Network. Since the categories are new, so in order to compare the results of the previous pipelines, we created a baseline classifier using simple rule-based string matching technique. The rule-based algorithm conditions on the occurrence of words such as "*janganlah, sholatlah, and so on*" to determine the category. The baseline method achieved F1-Score of 0.69, while the best F1-Score from the machine learning approach was 0.88, and it was produced by SVM model with the linear kernel.

1. Introduction

A hadith is a report that describes words, actions, or habits of Muslim Prophet, Muhammad [1]. There are many collections of hadith compiled by different scholars, some of them are more famous than the others. Hadith is the second reference to Islamic laws for Muslim after Qur'an. Usually, there are two types of content regarding a law, which are positive suggestion and negative suggestion of doing something. However, in hadith, there are also contents that do not fall under the two types, which will be categorized into information. The task of categorizing hadith into aforementioned types will benefit bigger system such as information retrieval or recommender system of hadith, because the system can suggest not only other hadith in the same topic, but also more specifically in the same type of either positive suggestion, negative suggestion, or just general information. Moreover, this kind of categorization is not a special task for hadith, but can be done also for other kinds of documents, for example, agreement documents.

One of the well-known hadith classification tasks is to decide if a hadith is authentic or not. This type of classification requires experts in the area of hadith science [2]. On the other hand, we try to position our hadith classification task into a common natural language text classification task, where the label of a text is determined only by its content without relying too much on the external knowledge. Also, we consider that common people are able to decide the appropriate label of a hadith given that they understand the language used by the hadith.

Our contribution through this paper is providing preliminary results on a new kind of text classification namely positive suggestion, negative suggestion, and information, especially text data from hadith translation in Bahasa.

2. Related Work

2.1. Hadith Authenticity

Usually, classification of hadith done by many Muslim scholars is to decide whether a hadith is authentic or unauthentic. There are common criteria in deciding the authenticity which are the reference to a particular authority, links of Isnad, number of reporters involved in each stage of Isnad, nature of the text and Isnad, and reliability and memory of the reporters [3]. To evaluate all of the criteria requires deep knowledge in the hadith science. Moreover, this is a very sensitive and important task, thus it should be done only by an expert in the hadith science, and it usually does not involve any computational method.

Nevertheless, there was a quite rare research to build an expert system for this type of classification, although the authors already stated that the purposes were to assist an expert in making the decision or to help non-expert to understand how the rules behind a decision [2]. Another similar research tried to focus on handling missing data in the hadith database and experimented with a decision tree and naive bayes as the classifiers [4]. On the other hand, our research goes to a different type of classification and is a bit less sensitive than deciding authenticity. Therefore, we can not relate our work very well to their research.

2.2. Text Classification

Text classification has been studied widely for decades. Joachims showed that SVM was superior to some other classification algorithms including Bayes, C.45, and KNN where tested on Reuters news dataset [5]. Besides SVM, Neural Network is another algorithm that performs good results on text classification [6, 7, 8]. Regarding text classification in Bahasa, there is previous research on lyric classification in Bahasa using Naive Bayes, which shows F1-Score of 0.89 [9]. Naradhipa and Purwarianti experimented with Naive Bayes, Maximum Entropy, and SVM to classify sentiment for product review in Bahasa, and the best accuracy of 81.6% was produced by SVM model [10]. Based on those previous results, we choose to experiment with SVM and Neural Network as the classifiers reported in this paper.

3. Data and Method

3.1. Hadith Dataset

Hadith dataset used in this work is the translation from Arabic to Bahasa, and they were collected from Imam Bukhari Collection, which is part of "Kitab Hadits 9 Imam" provided by Lidwa application[1]. The original dataset has no labels and still contain the *"isnad"* (list of scholars' names who reported the hadith). Data preparation was needed to separate and remove the *isnad* from the actual content of the hadith. In order to create a proper dataset for training and testing, each hadith was labeled manually by authors and colleagues[2]. There is a total of 1650 hadiths were labeled, 550 for each class of positive suggestion, negative suggestion, and information. Samples of hadith from each class are given in table 1.

Note that the keyword in the positive suggestion hadith is *"hendaklah"* (let him), while in negative suggestion is *"janganlah"* (do not). By identifying these keywords, we created a baseline classifier using rule-based string matching which will be elaborated in section 3.

[1] www.lidwa.com

[2] Aknowledgement for Andina, Eliza, Syair, and Fauzan for their contributions in labeling the hadiths

Table 1. Sample of hadith.

Class	Sample Hadith
Positive Suggestion	Jika seorang dari kalian memasuki masjid sedang imam sedang berkhuthbah atau dia telah keluar (kemudian masuk lagi) maka hendaklah dia shalat dua raka'at. English translation: *If one of you entered the mosque while the imam is preaching or he has come out (then come back) then let him pray two rak'ahs*
Negative Suggestion	Janganlah kalian berdusta terhadapku (atas namaku), karena barangsiapa berdusta terhadapku dia akan masuk neraka. English translation: *Do not tell lies about me (on my behalf), for whoever tells lies about me will enter Hell*
Information	Bukanlah disebut miskin orang yang bisa diatasi dengan satu atau dua suap makanan. Akan tetapi yang disebut miskin adalah orang yang tidak memiliki kecukupan namun dia menahan diri (malu) atau orang yang tidak meminta-minta secara mendesak. English translation: *The poor person is not the one who asks a morsel or two (of meals) from the others, but the poor is the one who has nothing and is ashamed to beg from others*

3.2. Method

Methods used in the experiment involve three main stages, preprocessing, feature extraction, and classification.

Preprocessing

As the common practice in text processing, the experiments began with preprocessing the dataset with one or more of these techniques: punctuation removal, case folding, stemming[3], and stopword removal.

Feature Extraction

The default feature extraction method used in our experiments was the combination of unigram and TF-IDF (Term Frequency-Inverse Document Frequency). But we also tried with different feature representations such as normal bag-of-word, bigram, and also trigram [11].

Baseline Classifier

To the best of our knowledge, the classes (i.e positive suggestion, negative suggestion, and information) and the dataset used in this work have not been studied before, so we do not have baseline results to compare with. Instead, we create our own baseline classification model using a simple rule-based and string matching. The rules are derived from a quick observation of the sample dataset, as described by Algorithm 1.

[3] https://github.com/sastrawi/sastrawi

if *one or more words in Negative Suggestion list appear in hadith* **then**
| classify as Negative Suggestion;
else if *one or more words in Positive Suggestion list appear in hadith* **then**
| classify as Positive Suggestion
else
| classify as Information
end

Algorithm 1: Rule-based classifier for the baseline

where the examples of words in each category list are given by Table 2.

Table 2. Example of words in Negative Suggestion and Positive Suggestion list.

Negative Suggestion	Positive Suggestion
jangan (do not)	hendaklah (let)
larang (forbid)	sholatlah (do pray)
tidak boleh (should not)	bersedekahlah (do give alm)
	Lakukanlah (do)
	Berpuasalah (do fasting)
	Jagalah (take care)

Support Vector Machine
Originally, Support Vector Machine (SVM) is a non-probabilistic classifier that only handles linear binary classification [12]. Since we have three classes, we have to build a multiclass SVM, in this case using one vs all method. First, we introduce a default SVM model called $\text{SVM}_{\text{default}}$, which uses Linear kernel, and C (trade-off parameter) value of 1. Later in the experiments, we tried various combinations of C value, Kernel method, and the kernel specific parameter.

Neural Network
As for Neural Network, our $\text{ANN}_{\text{default}}$ is configured with 10 hidden layers, and 10 neurons at each of the layers [13]. An alternative configuration we tried consisted of 20 hidden layers with 10 neurons at each of the layers.

4. Experiments and Results
Models were evaluated using 5-fold cross validation[14], so that at each fold 80% of the dataset was used for training and the remaining 20% for testing. The evaluation metric used throughout the experiments was macro-average F1-score [15].

4.1. Default models
In this experiment, we compared three classifiers on 5-fold cross validation settings. Preprocessing and feature extraction steps were not needed for baseline model. Meanwhile, for $\text{SVM}_{\text{default}}$ and $\text{NN}_{\text{default}}$, both preprocessing and feature extraction were set using combination of unigram and TF-IDF. Macro-average F1-score for the three models are presented in table 3.

Table 3. F1-Score for different classification models.

Model	F1-Score
Baseline	0.69
$SVM_{default}$	0.88
$ANN_{default}$	0.79

As we can see, our Baseline model just scored 0.69, which is far below the others. It means that the simple rule string matching classifier is not enough to predict the labels, and we need a more sophisticated classifier to do better. One of the examples where simple rules will not work is given in table 4.

Table 4. Missclassified hadith.

Bahasa	Barangsiapa berdusta terhadapku maka **hendaklah** ia persiapkan tempat duduknya di neraka.
English	*Whoever lies against me then **let him** prepare his seat in hell.*

If only we looked at the keyword, it would be classified as a positive suggestion, but the actual meaning of the hadith is a negative suggestion. This is a similar problem with sentiment analysis when processing sarcastic sentence[16].

4.2. Kernels

$SVM_{default}$ used Linear Kernel, so in this experiment, we want to know how different kernel influence the F1-Score of the SVM model. We tried Linear, Polynomial, and RBF kernels with various combination of degree (polynomial) and γ (rbf) values with fix C=1. In table 5, we reported the best result of each kernel.

Table 5. F1-Score for different SVM Kernels.

Kernel	F1-Score
Linear	0.8839
Polynomial	0.8779
RBF	0.8758

The best F1-score was produced by the linear kernel, not the more complex kernel like polynomial and RBF. The reason is probably that because of the number of features is a lot, so mapping the data point to a higher dimension will not help anymore [17].

4.3. Feature Representation
In this experiment, we tried different feature representation, namely unigram, bigram, and trigram which will be combined with TF-IDF weighting score.

Table 6. F1-Score for different feature representations.

Feature	F1-Score
Unigram	0.79
Bigram	0.70
Trigram	0.48

As shown in table 6, bigram and trigram representation did not help to improve what was achieved by unigram model.

4.4. Network Architecture
To investigate the potential improvement of Neural Network model, we construct a more complex architecture that contains 20 hidden layers and 10 neurons at each layers. The comparison with the NN with 10 hidden as default can be seen in table 7.

Table 7. F1-Score for different network architecture.

Architecture	F1-Score
10 hidden layers @ 10 neurons	0.79
20 hidden layers @ 10 neurons	0.85

With changing the architecture of NN, the performance changed significantly. In this case, an improvement of 0.06 point to the F1-Score. The need of more hidden layers can be related to the high input dimension of word features.

5. Conclusion
Classifying text into positive suggestion, negative suggestion, and information can be handled as normal text classification. In our experiments, SVM with linear kernel achieved the best F1-Score of 0.88. The classification problem is not as simple as we thought, as the simple rule-based string matching classifier did not perform well. The future work might be to investigate different feature extraction method, since suffixes (e.g lah in makanlah, sholatlah) play an important role in the meaning of the text, especially in positive suggestion class.

References
[1] Brown JA. Hadith: Muhammad's legacy in the medieval and modern world. Oneworld Publications; 2009 Jun 1.
[2] Bilal K, Mohsin S. Muhadith: A cloud based distributed expert system for classification of ahadith. InFrontiers of Information Technology (FIT), 2012 10th International Conference on 2012 Dec 17 (pp. 73-78). IEEE.
[3] Suhaib H. An introduction to the science of Hadith. London: Al-Quran Society, 1994.
[4] Aldhaln K, Zeki A, Zeki A, Alreshidi H. Improving knowledge extraction of Hadith classifier using decision tree algorithm. InInformation Retrieval & Knowledge Management (CAMP), 2012 International Conference on 2012 Mar 13 (pp. 148-152). IEEE.

[5] Joachims T. Text categorization with support vector machines: Learning with many relevant features. Machine learning: ECML-98. 1998:137-42.
[6] Aggarwal CC, Zhai C. A survey of text classification algorithms. Mining text data. 2012:163-222.
[7] Zhang ML, Zhou ZH. Multilabel neural networks with applications to functional genomics and text categorization. IEEE transactions on Knowledge and Data Engineering. 2006 Oct;18(10):1338-51.
[8] Sebastiani F. Machine learning in automated text categorization. ACM computing surveys (CSUR). 2002 Mar 1;34(1):1-47.
[9] Rifqi AA, Mubarok MS, Adiwijaya. Klasifikasi Topik pada Lirik Lagu dengan Metode Multinomial Naive Bayes. Indonesia Symposium on Computing (IndoSC) 2016. 2016.
[10] Naradhipa AR, Purwarianti A. Sentiment classification for Indonesian message in social media. InCloud Computing and Social Networking (ICCCSN), 2012 International Conference on 2012 Apr 26 (pp. 1-5). IEEE.
[11] Jurafsky D, Martin JH. Speech and language processing. London: Pearson; 2014 Dec 30.
[12] Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995 Sep 1;20(3):273-97.
[13] Demuth HB, Beale MH, De Jess O, Hagan MT. Neural network design. Martin Hagan; 2014 Sep 1.
[14] Kohavi R. "A study of cross-validation and bootstrap for accuracy estimation and model selection." Ijcai. Vol. 14. No. 2. 1995.
[15] Yang Y. An evaluation of statistical approaches to text categorization. Information retrieval. 1999 Apr 1;1(1):69-90.
[16] Maynard D, Mark AG. Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis. LREC. 2014.
[17] Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification.