Eurasian Journal of Analytical Chemistry
ISSN: 1306-3057 OPEN ACCESS 2018 13 (6): 186-191

Classification of Hadith Levels Using Data and Text Mining Techniques

Bustami, Muhammad Fikry

Received 14 September 2018 ▪ Revised 23 October 2018 ▪ Accepted 24 November 2018

Abstract: There is no definite information among ulama, about the beginning of the Prophet's hadith forgery, but this problem has spread and responded to the community. The purpose of the hadith counterfeiters are various motives and motivations, the factors that encourage them to falsify the hadith are to defend certain interests: defending political interests, defending theology, defending fiqh madzhab, attracting people who hear their stories, to dignify others, encourage others are more persistent in worshiping and destroying Islam. Determining the level of hadith requires a long process because we have to read the entire hadith and know the perawi and sanad. This problem requires a solution to overcome it. Through this research, search and analysis of the model was carried out using the Naïve Bayes (NB) algorithm and the Decision Tree algorithm (C4.5). Evaluation is done by comparing two algorithms. Based on the results of the study found that the Decision Tree Algorithm (C4.5) has a higher accuracy rate of 7.81% of the Naïve Bayes Classifier Algorithm.

Keywords: Naïve Bayes, Decision Tree, C4.5, Hadith.

INTRODUCTION

Hadith as a source of law in the Islamic religion has a second position at the level of the source of law under the Qur'aan. Hadith literally means words or conversation. In Islamic terminology, the term hadith means to report/record a statement and behavior of the Prophet (ﷺ). But at this time the word hadith experienced an expansion of meaning so that it was synonymous with the Sunnah, it could mean all words, deeds, provisions, and approval of the Prophet (ﷺ)which were made as provisions or laws. The word hadith itself is not an infinitive word, then the word is a noun included in the category of hadith is atsar, which is something that is leaning on the companions of the Prophet (ﷺ) and also taqrir, namely the state of the Prophet (ﷺ) who silenced, does not hold objections or approves what the friends have done or said before him. Throughout his life, Rasulullah was always there to answer various questions and problems faced by the people at any time. Therefore, no person has made a record regarding his statement. In fact, Rasulullah forbade his people to write something if it was not Al - Qur'an. Because he feared that people will mix Al-Quran with its words besides revelation. As a result, the greatest pressure regarding writing is placed on the recording of the Qur'aanic verses. However, there are many authentic narratives collected by the Hadith scholars who prove that the Hadith is recorded in writing even during the lifetime of the Prophet (ﷺ).

To be able to know a hadith into a certain level is not easy, it is necessary to read the whole of the hadith then find out about the narrators and sanad of the hadith. The level of the hadith is *saheeh*, *dhaif*, *dhaifjiddan*, *bathil*, *munkar*, *maudhu*. This is what makes it difficult to determine a hadith level, while many people cannot understand the narrators and sanad. Problems often occur with people who want to know the truth of the information from the hadith that is read. Based on these problems an alternative can be taken by utilizing data mining techniques by comparing 2 methods for the classification of hadith levels using the Decision Tree (C4.5) and Naïve Bayes methods. With hope, after processing with data mining can help provide information about the truth of the hadith. The research conducted is by tracing the words in a text from the hadith, then comparing the accuracy of the information from using Decision Tree (C4.5) and Naïve Bayes algorithms.

Bustami, Doctoral Student of Mathematics and Applied Science, Syiah Kuala University.
Muhammad Fikry, Department of Informatics, Faculty of Engineering, Universitas Malikussaleh.

RELATED WORK

In this section, we present important studies that have been carried out in hadith data mining analysis, decision tree (C4.5) and naïve bayes algorithms. The reason for this study was due to the lack of complexity of analysis in determining the level of hadith.

Naïve Bayes

The Naïve Bayes algorithm is a classification method using probability and statistical methods. The Naïve Bayes algorithm predicts opportunities based on existing data so that it is known as the Bayes Theorem. Bayes's theorem is used to calculate posterior probabilities, P (c | x), from P (c), P (x), and P (x | c). The classifier carried out by Naïve Bayes is by analyzing the effect of the (x) predictor value on a particular class (c) does not depend on other predictor values. This analogy is also called the conditional dependency class.

Likelihood Class Prior Probability

$$P(c \mid x)=\frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c)\times P(x_2 \mid c)\times \cdots \times P(x_n \mid c)\times P(c)$$

- P (c | x) is a posterior probability (target) which is given a predictor value (attribute).
- P(c) is the prior probability of class.
- P(x|c) is the probability of predictor probability which is then given a class.
- P(x) is the prior probability of predictor.

Numerical variables must be transformed into existing categories before it forms its own frequency table. The other option is to distribute numeric variables to estimate the frequency. An example is to assume a normal distribution for numeric variables. The probability density function is for a normal distribution which is then defined by two parameters (mean and standard deviation).

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \boxed{\text{Mean}}$$

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2\right]^{0.5} \qquad \boxed{\text{Standard deviation}}$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \boxed{\text{Normal distribution}}$$

Profit information as the amount of information the district by each attribute can provide an explanation of how predictor values affect the probability of class.

$$log_2 P(c|x) - log_2 P(c)$$

Decision Tree (C4.5)

C4.5 data mining algorithm is one of the algorithms used to classify or segment or classify and be predictive. Selection of good attributes is an attribute that allows getting the smallest decision tree size. Or attributes that can separate objects according to their class. Heuristically the attribute selected is the attribute that produces the most "purest" node. The size of the purity is expressed by the level of impurity, and to calculate it, can be done using the concept of Entropy, Entropy states the impurity of a collection of objects.

$$Entropi\ (S) = \sum_{j=1}^{k} -p_j \log_2 p_j$$

- S is a case set
- k is the number of partitions S
- pj is the probability obtained from Sum (Yes) divided by the Total Case

Information gain is the most popular criterion for attribute selection. The C4.5 algorithm is the development of the ID3 algorithm. Because of this development, the C4.5 algorithm has the same basic working principles as the ID3 algorithm. It's just that in the C4.5 algorithm the attribute selection is done by using Gain Ratio with the formula:

$$gain\ ratio(a) = \frac{gain(a)}{split(a)}$$

- a = attribute
- gain (a) = information gain on attribute a
- Split (a) = split information in attribute a

Attributes with the highest Gain Ratio value are selected as test attributes for vertices. With gain is information gain. This approach applies normalization to information gain by using what is called split information. SplitInfo states entropy or potential information with the formula:

$$SplitInfo(S,A) = -\sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S}$$

- S = space (data) sample used for training
- A = attribute
- Si = number of samples for attributes i

where Xi represents the i subset in sample X.

$$Gain\ (A) = Entropi\ (S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} \times Entropi(S_i)$$

S = space (data) sample used for training.

A = attribute.

|Si| = the number of samples for the value of V.

|S| = number of all data samples.

Entropy (Si) = entropy for samples that have a value of i

Data and Text Mining

The difference between data mining, association rule mining, and text mining is that in text mining, patterns are extracted from natural language text, but data mining and association rule mining patterns are extracted from databases. The steps in the mining process are (Chiwara):

- Text: This represents the given target document for mining in text format.
- Text processing: This step concern to text clean up, format, tokenize and others.
- Text transformation (attribute generation): Generates attributes from text that has been processed based on the text provided
- Attribute selection: Select attributes for mining data because not all attributes produced will be suitable for mining
- Data Mining (Pattern discovery): Mine the selected attribute and then extract it according to the desired pattern.
- Interpretation and evaluation: This is about what you are looking for in the next step, i.e. terminate, results that are perfect for the application that you want and so on.

Many people do not know the difference between data mining and knowledge discovery, others think that data mining is the main stage of the Knowledge Discovery in Database (KDD) process. KDD is based on the whole process that extracts useful knowledge from the amount of data available. Including evaluations to make decisions about what is a requirement to become a knowledge. Whereas data mining refers to the application of algorithms to extract patterns from the data. KDD's steps are as follows,

- Data Cleaning: Removing noise or outliers which interferes with data retrievaland inconsistent.
- Data Integration: Combining data from considerable data sources
- Data Selection: Relevant data by processing data from existing databases.
- Data Transformation: Transforming or entering data into desired forms for data mining by conducting operations or aggregation.
- Data Mining: Applying intelligent methods to extract data patterns that it have.
- Pattern Evaluation: Evaluating data patternswich exists.

- Knowledge Presentation: Representing knowledge collected.

RESULTS AND DISCUSSIONS

In this study, an application program was made using PHP's Promrograming language with the SQL database. In the application created, each algorithm has been included in the application.

In the Naïve Bayes experiment, it was carried out by entering the search keywords for the available hadith with the number of datasets as many as 148 hadiths. The following is an example of the hadith used as one of the research datasets.

> Matan:
>
> تُقَدِّمُوْاسُفَهَاءَكُمْفِيصَلَاتِكُمْوَلَاعَلَىجَنَائِزِكُمْفَإِنَّهُمْوَفْدُكُمْإِلَىرَبِّكُلَا) : عَنْعَلِيِّرَضِيَاللَّهُعَنْهُقَالَرَسُوْلَاللَّهِصَلَّاللَّهِعَلَيْهِوَسَلَّ
> مْ)
>
> Translate in Indonesian:
>
> Dari Ali Radhiayallahu ‘Anhu, Rasulullah Shallallahu ‘Alaihi Wasallam bersabda: “Janganlah kamu mendahulukan orang-orang yang bodoh dari kamu (untuk menjadi iman) dalam shalat, juga janganlah mendahulukan untuk shalat atas jenazahmu sesungguhnya mereka adalah utusanmu untuk Rabbmu.”

The experimental results of the database that have been obtained using the Naive Bayes algorithm are as follows:

Table 1: The obtained results from Naïve Bayes experiments

Words	The Accuracy of Averages
Rasullullah	88,32%
Shalat	78,31%
Air	69,54%
Allah	94,50%
Telah menceritakan kepada kami	60,20%
berbicara dusta	55,02%
alhamdulillah	52,55%
Beruntunglah orang yang diam	38,48%
Islam	57,19%
seorang hamba	55,02%

From the Naïve Bayes experiment table on the dataset, the highest accuracy is 94.5%, the lowest accuracy level is 38.48%, and the average accuracy is 64.91%.

Results of the Decision Tree Experiment (C4.5)

In the C4.5 algorithm experiment, it is done by entering the same keyword as the Naive Bayes algorithm experiment. The following are the experimental results of the datasets that have been obtained using the C4.5 algorithm.

Table 2: The obtained results from C4.5 experiments

Words	The Accuracy of Averages
Rasullullah	89,84%
Shalat	83,72%
Air	75,14%
Allah	95,01%
Telah menceritakan kepada kami	72,02%
berbicara dusta	69,57%
alhamdulillah	62%
Beruntunglah orang yang diam	53,03%
Islam	57,36%
seorang hamba	69,57%

From the C4.5 experiment table on the dataset, the highest accuracy is 95.01%, the lowest accuracy level is 53.03%, and the average accuracy is 72.73%.

Validation

Validation is an action that proves that a process/method can provide consistent results in accordance with established specifications and is well documented.

After experimenting on the Naive Bayes and C4.5 algorithms, the following algorithm is obtained:

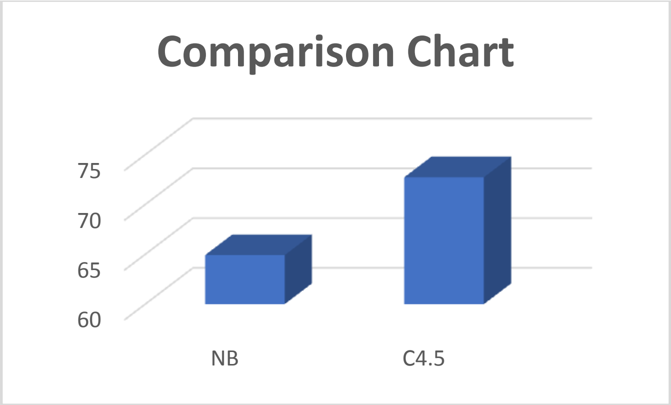

Fig .1: Comparison Chart

CONCLUSION

The experimental results of the Decision Tree algorithm (C4.5), resulted in an accuracy rate of 7.81% greater than that of Naïve Bayes. From the results of the study, it can be concluded that the C4.5 algorithm has a better level of accuracy compared to Naïve Bayes for the classification of hadith levels. However, it is recommended for further research to experiment with other models so that a better level of accuracy can be obtained.

REFERENCES

[1] Brahimi, B., Touahria, M., & Tari, A. (2016). Data and Text Mining Techniques for Classifying Arabic Tweet Polarity. *Journal of Digital Information Management*, *14*(1).

[2] Falotico, R., Liberati, C., & Zappa, P. (2015). Identifying Oncological Patient Information Needs to Improve e‐Health Communication: a preliminary text‐mining analysis. *Quality and Reliability Engineering International*, *31*(7), 1115-1126.

[3] Fikry, M. (2016). Rancangan basis data kependudukan berdasarkan aspek-aspek kualitas schema database. *TECHSI-Jurnal Teknik Informatika*, *8*(2).

[4] Fikry, M. (2016). Aplikasi Java Kriptografi Menggunakan Algoritma Vigenere. *Journal Techsi*, *8*(1), 1-9.

[5] Khashfeh, M., Mahmoud, M. a., & Ahmad, M. S. (2018). An analysis of text mining factors enhancing the identification of relevant studies. *Journal of Theoretical & Applied Information Technology*, *96*(12).

[6] Kim, Y., & Jeong, S. R. (2018). Competitive intelligence in Korean Ramen Market using Text Mining and Sentiment Analysis. *인터넷정보학회논문지*, *19*(1), 155-166..

[7] Kinne, J., & Axenbeck, J. (2018). Web mining of firm websites: A framework for web scraping and a pilot study for Germany. *ZEW-Centre for European Economic Research Discussion Paper*.

[8] Philips, A. A. B. (2005). Asal-Usul dan Perkembangan Fiqh: Analisis Historis Atas Mazhab, Doktrin dan Kontribusi, terj. *M. Fauzi Arifin, Bandung: Nuansa*.

[9] Philips, Abu Ameenah Bilal. Usool Hadith: The Methodology of Hadith Evaluation. *IslamKotob*, 1990.

[10] Rosandy, T. (2017). Perbandingan Metode Naive Bayes Classifier Dengan Metode Decision Tree (C4. 5) Untuk Menganalisa Kelancaran Pembiayaan (Study Kasus: Kspps/Bmt Al-fadhila. *Jurnal Teknologi Informasi Magister*, *2*(01), 52-62.

[11] Saloot, M. A., Idris, N., Mahmud, R., Ja'afar, S., Thorleuchter, D., & Gani, A. (2016). Hadith data mining and classification: a comparative analysis. *Artificial Intelligence Review*, *46*(1), 113-128.

[12] Sarma, G. (2017). Scientific literature text mining and the case for Open Access. *The Journal of Open Engineering*.

[13] Somantri, O. (2017). Text Mining Untuk Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes (NB). *Jurnal Telematika*, *12*(1), 7-12.

[14] Singh, R., Rani, A., Kumar, P., Singh, C., Shukla, G., & Kumar, A. Hemicellulolytic Activity in the Crop Residues. *International journal of pharmacy research & technology*, 7(1), 18-20.

[15] Mathumathi, K.M. and Senthilprakash, K. (2017) energy sentient qos implemented node-disjoint multipath routing protocolfor manet. *International journal of communication and computer technologies*, 5 (2), 67-75.

[16] Surendar, A. (2018). Role of Microbiology in the Pharmaceutical & Medical Device. *International Journal of Pharmaceutical Research*, *10*(3).