Klasifikasi Text Mining untuk Terjemahan Ayat-Ayat Al-Qur'an menggunakan Metode Klasifikasi Naive Bayes

Text Mining Classification for Translation of Al-Qur'an Verses using the Naive Bayes Classifier

[1]Nadya Hilwah, [2]Abdul Kudus, [3]Siti Sunendiari

[1,2,3]Prodi Statistika, FakultasMatematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Jl. Ranggamalela No.1 Bandung 40116

email: [1]nadyahilwah@gmail.com, [2]akudus69@yahoo.com, [3]sunen_diari@yahoo.com

Abstract. The Qur'an is a document and a shield for guidance especially for Muslims (Muslim). In Qur'anic the studying, people have to knowledge at least about the classification in the Qur'an. Which have the function in studying further of a verse and facilitate in searching for verses related to something else or beetwen verses relationship. One method that can be used for the classification of Qur'an verses is Naive Bayes Classifier method. This is one of the statistical methods that used to perform classification analysis. This paper discusses about formation of document classification models using it. Verse classification is based on the wordlist in the verse. Then class labeling is based on selected words. In this study, the document used is document translation of the verses of Qur'an. The Naive Bayes model is applied to data testing, which is a document of Qur'anic verses that do not yet have a class label. The result of this research are: (1) There are 198 wordlist from result of preproccessing which is used to process formation of Naive Bayes classification model. (2) The prediction results of the new document using 15 models that have been established, there are 47 verses that do not belong to the 15 predetermined categories, while the other 707 verses have been determined in that categories.

Keywords : Qur'an, Classification, Naive Bayes.

Abstrak. Al-Qur'an adalah sebuah dokumen dan sebuah perisai untuk umat manusia khususnya untuk umat yang beragama Islam (Muslim). Dalam mempelajari Ilmu Al-Qur'an seseorang minimal harus mengetahui tentang pengklasifikasian yang terdapat dalam Al-Qur'an. Dimana pengklasifikasian tersebut dapat berfungsi untuk pengkajian lanjut mengenai sebuah ayat dan memudahkan dalam mencari ayat-ayat yang berhubungan dengan sesuatu atau dalam mencari hubungan ayat satu dengan ayat lainnya. Salah satu metode yang dapat digunakan untuk pengklasifikasian ayat-ayat Al-Qur'an yaitu metode klasifikasi Naive Bayes. Naive Bayes adalah salah satu metode statistika yang digunakan untuk melakukan analisis klasifikasi. Makalah ini membahas cara pembentukan model klasifikasi dokumen dengan menggunakan metode klasifikasi Naive Bayes. Pengklasifikasian suatu ayat berdasarkan pada kata-kata penting dalam ayat tersebut. Dan pelabelan kelas didasarkan pada kata-kata yang terseleksi. Dalam penelitian ini dokumen yang digunakan adalah dokumen terjemahan ayat-ayat Al-Qur'an. Model Naive Bayes diterapkan pada data testing, yaitu dokumen ayat-ayat Al-Qur'an yang belum memiliki label kelas. Hasil dari penelitian ini adalah: (1) Terdapat 198 wordlist dari hasil preproccessing yang digunakan untuk proses pembentukan model klasifikasi Naive Bayes. (2) Hasil prediksi dari dokumen baru menggunakan 15 model yang telah dibentuk, terdapat 47 ayat yang tidak termasuk ke dalam 15 kategori yang telah ditentukan, sedangkan 707 ayat lainnya termasuk ke dalam beberapa kategori yang telah ditentukan.

Kata Kunci : Al-Qur'an, Klasifikasi, Naive Bayes.

A. Pendahuluan

Al-Qur'an adalah sebuah dokumen dan sebuah perisai untuk umat manusia khususnya untuk umat yang beragama Islam (Muslim). Ia sendiri menamakan dirinya "petunjuk bagi umat manusia" (*Hudan li al-Nas*).

Dalam mempelajari Ilmu Al-Qur'an, secara otomatis seseorang minimal harus mengetahui tentang pengklasifikasian ayat-ayat yang terdapat dalam Al-Qur'an. Pentingnya seseorang mengetahui pengklasifikasian dalam Al-Qur'an menjadi suatu keharusan. Dimana pengklasifikasian tersebut dapat berfungsi untuk mengkaji lebih lanjut mengenai sebuah ayat dan memudahkan dalam mencari ayat-ayat yang berhubungan, sepoerti hubungan ayat satu dengan ayat lainnya.

Seiring perkembangan zaman, beberapa penerbit buku yang menerbitkan Al-Qur'an dan Terjemahan melampirkan jenis kategori ayat-ayat Al-Qur'an ke dalam beberapa kelas. Dalam Al-Qur'an dan Terjemahan yang diterbitkan oleh penerbit Cordova, melampirkan bahwa dalam Al-Qur'an terdapat 15 klasifikasi ayat-ayat Al-Qur'an. Namun tidak semua ayat memiliki label kelas, ada beberapa ayat yang belum memiliki kelas. Sehingga ayat-ayat yang belum memiliki label kelas tersebut dapat diprediksi dangan menggali informasi dari ayat yang sudah memiliki label kelasnya.

Salah satu cara untuk memprediksi klasifikasi kelas ayat-ayat Al-Qur'an menurut Ilmu Statistika yaitu dengan cara membuat suatu model. Pemodelan ini dapat menerapkan salah satu teknik *data mining* (penambangan data) yaitu klasifikasi. Metode klasifikasi yang dapat digunakan untuk memprediksi klasifikasi kelas ayat-ayat Al-Qur'an adalah metode klasifikasi *Naive Bayes*. Analisis ini merupakan salah satu metode dalam data mining yang bertujuan untuk memprediksi ayat-ayat Al-Qur'an yang belum memiliki label kelas.

Berdasarkan uraian dari latar belakang, maka masalah yang dapat diidentifikasi adalah bagaimana cara untuk memprediksi ayat-ayat Al-Qur'an yang belum memiliki label kelas berdasarkan metode klasifikasi Naive Bayes. Selanjutnya, tujuan yang ingin dicapai dalam penelitian ini antara lain.

1. Membentuk model klasifikasi kelas ayat-ayat Al-Qur'an dengan menggunakan metode klasifikasi *Naive Bayes*.
2. Memprediksi kelas ayat-ayat Al-Qur'an yang belum diketahui label kelasnya dengan menggunakan model klasifikasi *Naive Bayes*.

B. Landasan Teori

Data mining merupakan prinsip dasar dalam mengurutkan data dalam jumlah yang sangat banyak dan mengambil informasi – informasi yang berkaitan dengan apa yang diperlukan seperti apa yang biasa dilakukan oleh seorang analis. Dengan bertambah banyaknya jumlah data yang ada dalam model, maka peran analis untuk menganalisa data secara manual perlu digantikan dengan aplikasi yang berbasis komputer yang dapat menganalisa data secara otomatis menggunakan alat yang lebih kompleks dan canggih (Rully, 2009). *Data Mining* digunakan untuk menemukan "pengetahuan" penting dari database besar (Kudus, 2008).

Text mining memiliki definisi menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen, dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antardokumen (Maria, 2013).

Tujuan dari *text mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Jadi, sumber data yang digunakan pada *text mining* adalah kumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi terstruktur. Adapun tugas khusus dari text mining antara lain yaitu pengkategorisasian teks (*text categorization*) dan pengelompokan teks (*text clustering*).

Berdasarkan ketidakteraturan struktur data teks, maka proses *text mining* memerlukan beberapa tahap awal yang pada intinya adalah mempersiapkan agar teks dapat diubah menjadi lebih terstruktur. Adapun tahapan yang dilakukan secara umum dalam text mining terdapat empat proses pokok, yaitu:

1. Text Preprocessing

Tahap ini bertujuan untuk mempersiapkan teks menjadi data yang akan mengalami pengolahan lebih lanjut. Operasi yang dapat dilakukan pada tahap ini meliputi:

a. *Transform case*
   Proses Transform case dilakukan untuk mengubah kapitalisasi karakter (huruf) menjadi kecil untuk semua kata atau huruf.
b. *Tokenization*
   Proses tokenization dilakukan untuk membuang semua karakter yang tidak dibutuhkan.
c. *Filter Token*
   Proses filter token merupakan proses pembuangan kata-kata yang kurang dari batas minimal yang sudah ditentukan oleh peneliti.
d. *Stopwords*
   Stopwords merupakan kumpulan daftar kata-kata yang kemungkinan besar tidak akan memberikan pengaruh prediksi, seperti imbuhan dan kata ganti. Kata yang termasuk stopwords akan dibuang, karena kata tersebut tidak merepresentasikan isi dokumen walaupun sering muncul.
e. *Stemming*
   Tahap stemming merupakan proses untuk mengubah semua kata yang telah dipilih pada proses tokenization menjadi kata yang berupa kata dasar.
f. Setelah semua data yang telah diproses pada tahap *preproccessing* sebelumnya, selanjutnya akan diproses untuk dilakukan transformasi yang menghasilkan daftar kata-kata penting (*wordlist*) agar data teks menjadi terstruktur. Setiap kata pada wordlist tersebut harus diberi nilai bobot agar bisa dihitung tingkat kepentingan setiap kata. Dalam penelitian ini, pembobotan kata akan menggunakan TF-IDF, yang merupakan penggabungan dengan cara mengalikan *term frequency* (tf) dengan nilai *invers document frequency* (idf). Dengan demikian rumus umum untuk TF-IDF adalah sebagai berikut:

$$w(d.t) = tf(d.t) \times idf_t$$
$$= tf(d.t) \times \log(D / df_t)$$

Keterangan:
$w(d.t)$ = bobot kata(t) terhadap dokumen di
$tf(d.t)$ = jumlah kemunculan kata(t) dalam dokumen di
D = jumlah dokumen pada dataset
df_t = jumlah dokumen yang mengandung kata(t)

2. Validasi dan Evaluasi Hasil

Dalam penelitian ini, valaidasi dan Evaluasi Hasil menggunakan metode klasifikasi *Naive Bayes.*

Klasifikasi *Naive Bayes* adalah suatu klasifikasi berpeluang sederhana berdasarkan aplikasi teorema *Bayes* dengan asumsi antar variabel penjelas saling bebas (independen). Dalam hal ini, diasumsikan bahwa kehadiran atau ketiadaan dari suatu kejadian tertentu dari suatu kelompok tidak berhubungan dengan kehadiran atau ketiadaan dari kejadian lainnya. *Naive Bayes* dapat digunakan untuk berbagai macam keperluan antara lain untuk klasifikasi dokumen, deteksi spam atau *spam filtering*, dan masalah klasifikasi lainnya. Dalam hal ini lebih disorot mengenai penggunaan teorema *Naive Bayes* untuk *spam filtering* (Nusantara, 2015).

Secara umum persamaan teorema Bayes dapat ditulis sebagai berikut:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Dengan menerapkan teorema *Bayes* pada persamaan di atas, persamaan klasifikasi *Naive Bayes* dapat dinyatakan sebagai berikut:

$$P(Y = C_k \mid x_1, ..., x_n) = \frac{P(x_1, ..., x_n \mid Y = C_k) P(Y = C_k)}{P(x_1, ..., x_n)}$$

Pada persamaan di atas, $P(Y = C_k \mid x_1, ..., x_n)$ adalah peluang kategori Ck dari semua atribut yang berada dalam dokumen tertentu. $P(Y = C_k)$ adalah peluang dari kategori tertentu yang dibandingkan dengan kategori yang dianalisis lainnya. $P(x_1, ..., x_n)$ adalah peluang dari dokumen tertentu secara spesifik. Karena nilai $P(x_1, ..., x_n)$ untuk semua kategori besarnya sama, maka nilainya dapat diabaikan.

Selanjutnya dengan menggunakan *Maximum A Posteriori* (MAP) akan dicari nilai probabilitas untuk masing-masing kelas, dengan persamaan sebagai berikut:

$$C_{map} = \arg\max P(Y = C_k) \prod_{i=1}^{n} P(x_i \mid Y = C_k)$$

Sebelum menentukan di kelas mana dokumen baru akan diklasifikasikan, hitung terlebih dahulu *Prior Probability* $P(Y = C_k) = \frac{N_k}{N}$, dan *Likelihood* $P(x_i \mid Y = C_k) = \frac{n_t + 1}{n + |V|}$.

Keterangan:

N_k = jumlah dokumen setiap kategori k

N = jumlah dokumen dari semua kategori

n_t = jumlah kemunculan kata t dalam dokumen yang berkategori Ck

n = jumlah seluruh kata dalam dokumen berkategori Ck

$|V|$ = jumlah semua kosakata dari semua kategori

Untuk menghitung keakuratan model, dalam penelitian ini menggunakan *Confusion Matrix*. Nilai akurasi *confusion matrix* diperoleh dengan persamaan sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Nilai *true positive* (TP) dan *true negative* (TN) adalah hasil klasifikasi yang benar. Nilai *false positive* (FP) adalah nilai dimana hasilnya diprediksi sebagai *class-1* namun sebenarnya merupakan *class-2*. Sedangkan *false negative* (FN) adalah nilai dimana prediksi mengklasifikasikan sebagai *class-2* namun faktanya termasuk dalam klasifikasi *class-1*.

C. Hasil Penelitian dan Pembahasan

Ayat-ayat Al-Qur'an yang sudah memiliki label kelas sebanyak 5482 ayat, dan sisanya sebanyak 754 ayat belum memiliki label kelas. Artinya terdapat 88% ayat yang sudah memiliki label kelas, dan 12% ayat yang belum memiliki label kelas.

Adapun dari 88% ayat yang sudah memiliki label kelas akan dijadikan *data training*, dan 12% ayat yang belum memilki label kelas akan dijadikan *data testing*. Dari 88% data training tersebut akan dibagi menjadi 15 dataset dengan 15 kategori yang sudah ditentukan. Pada setiap dataset memiliki dua label kelas, yaitu kelas kategori dataset tersebut dan kelas bukan kategori dataset tersebut.

Preprocessing data pada penelitian ini menggunakan *software* STATISTICA. Pada tahap ini *preprocessing* data akan melalui beberapa tahapan, yaitu: tokenizing,

stopword, *filtering*, *stemming*, dan *token weightening* TF-IDF.

Setelah melakukan *preprocessing*, terdapat 198 daftar kata penting (*wordlist*) yang menjadi atribut untuk membentuk model *Naive Bayes.*

Berikut merupakan Tabel *Prior Probability* dari 15 dataset:

Tabel 1. Nilai Prior Probability

Kategori	P(Ya)	P(Tidak)
Rukun Islam	0.62	0.38
Iman	0.42	0.58
Al-Qur'an	0.1	0.9
Ilmu	0.1	0.9
Amal	0.13	0.87
Dakwah	0.04	0.96
Jihad	0.07	0.93
Manusia & Humas	0.11	0.89
Akhlak	0.14	0.86
Harta	0.95	0.05
Hukum	0.04	0.96
Negara & Masyarakat	0.01	0.99
Pertanian & Perdagangan	0.01	0.99
Sejarah & Kisah-Kisah	0.12	0.88
Agama-Agama	0.18	0.82

Berikut merupakan Tabel *Likelihood* untuk dataset kategori Rukun Islam:

Tabel 2. Nilai *Likelihood* kategori Rukun Islam

No	Term	nt (Yes)	nt (No)	Rukun Islam	
				P(xi|Ya)	P(xi|Tidak)
1	abid	49	40	0.00200	0.00328
2	abod	44	25	0.00180	0.00208
3	account	54	24	0.00220	0.00200
⋮					
198	would	204	92	0.00819	0.00744
Total (n)		24830	12309		

Untuk nilai $\mathbf{P(x_i|Ya,Tidak)}$ diperoleh dari persamaan *Likelihood*. Misal nilai $\mathbf{P(x_i|Ya)}$ pada kolom pertama diperoleh dari perhitungan [(49-1) / (24830-198)] = 0.00200. Begitupun dengan nilai *Likelihood* lainnya.

Nilai *prior probability* dan nilai *likelihood* di atas dapat digunakan untuk mencari nilai *posterior* atau nilai prediksi kelas dari suatu dokumen. Dalam pembentukan model pada tahap ini akan menghasilkan kelas prediksi untuk mengetahui keakuratan model. Misalkan untuk mengetahui prediksi kelas suatu dokumen pada dataset Rukun Islam yang berisi kata '*abid'* dan '*would'* dapat dihitung dengan *posterior probability* sebagai berikut:

P(Ya|X) = 0.62 [0.00200 * 0.00819]
= 0,00001

P(Tidak|X) = 0.38 [0.00328 * 0.00744]
= 0,000009

Dari perhitungan posterior di atas, menunjukkan bahwa nilai probabilitas yang paling besar berada pada kelas YA (Rukun Islam), sehingga dokumen tersebut diprediksi masuk ke dalam kelas Rukun Islam. Artinya dokumen tersebut berhasil diklasifikasikan dengan benar.

Nilai akurasi dari 15 dataset yang telah dibentuk adalah sebagai berikut:

Tabel 3. Nilai Akurasi Model

Kategori	Nilai Akurasi
Rukun Islam	62.0576%
Iman	57.7344%
Al-Qur'an	84.6771%
Ilmu	85.3338%
Amal	78.1649%
Dakwah	89.8212%
Jihad	81.1383%
Manusia & Humas	82.5611%
Akhlak	82.0139%
Kategori	**Nilai Akurasi**
Harta	89.4929%
Hukum	86.9208%
Negara & Masyarakat	93.7979%
Pertanian & Perdagangan	97.7381%
Sejarah & Kisah-Kisah	86.1%
Agama-Agama	78.6392%

Proses selanjutnya yaitu penerapan model pada data baru, yaitu ayat-ayat Al-Qur'an yang belum memiliki label kelas. Setelah dilakukan prediksi menggunakan model klasifikasi *Naive Bayes*, dari 754 ayat yang belum memiliki label kelas terdapat 47 ayat yang tidak termasuk ke dalam 15 kategori yang telah ditentukan, sedangkan 707 ayat lainnya termasuk ke dalam beberapa kategori. Hal tersebut terjadi karena dari pengujian terhadap 15 model, nilai *posterior* terbesar terdapat pada bukan kategori dari 15 dataset yang telah ditentukan. Berikut daftar surat-surat dan ayat-ayat Al-Qur'an yang tidak termasuk ke dalam 15 kategori yang sudah ditentukan:

Tabel 4. Surat dan Ayat yang tidak termasuk ke dalam 15 kategori

No	Surah ke-	Ayat Ke-
1	12	3
2		35
3		93
4	14	49
5	20	23
6		26
7		28
8		29
9		33
10		89
11		91
12		94
13	21	63
14	23	31
15	26	142
16		162
No	**Surah**	**Ayat Ke-**
	ke-	
17	27	55
18	29	28
19	34	20
20	37	26
21		77
22		98
23		175
24	38	36
25	43	56
26	51	33
27	54	12
28		13
29	56	73
30	68	5
31	69	18
32	71	6
No	**Surah ke-**	**Ayat Ke-**
33	75	22
34	77	12
35		19
36		28
37		40
38		45
39	78	32
40	79	46
41	80	38
42	82	17
43		18
44	84	7
45	93	1
46	100	3
47	105	5

D. Kesimpulan dan Saran

Kesimpulan

Naive Bayes merupakan sebuah teknik klasifikasi probabilistik yang berdasarkan teorema *Bayes* dengan asumsi variabel penjelas saling bebas. Pada penelitian ini, variabel penjelas merupakan suatu dokumen atau ayat. Dengan hal ini, ada atau tidak adanya suatu ayat tertentu, tidak akan berpengaruh terhadap ayat lainnya.

Berdasarkan hasil pengujian yang telah dilakukan pada bab sebelumnya, maka dapat ditarik kesimpulan sebagai berikut:

1. Hasil preprocessing dengan menggunakan pembobotan kata TF-IDF menghasilkan daftar kata 198 kata penting (wordlist).
2. Dari 6236 ayat yang ada dalam Al-Qur'an, terdapat 88% ayat dalam Al-Qur'an yang sudah memiliki label kelas, dan 12% ayat dalam Al-Qur'an yang belum memilik label kelas.
3. Dataset yang menghasilkan model klasifikasi dengan nilai akurasi tertinggi adalah dataset pada kategori Pertanian dan Perdagangan dengan nilai akurasi sebesar 97.7381%. Sedangkan dataset yang menghasilkan model klasifikasi dengan nilai akurasi terendah adalah dataset pada kategori Iman dengan nilai akurasi sebesar 57.7344%.
4. Hasil prediksi dari dokumen baru, terdapat 47 ayat yang tidak termasuk ke dalam 15 kategori yang telah ditentukan, sedangkan 707 ayat lainnya termasuk ke dalam beberapa kategori.

Saran

Dari hasil prediksi klasifikasi menggunakan model *Naive Bayes*, masih terdapat ayat-ayat yang tidak termasuk ke dalam 15 kategori yang sudah ditentukan. Maka dari itu, penulis menyarankan agar ada penambahan kategori untuk ayat-ayat yang belum terklasifikasi. Baik dengan menggunakan metode pengkajian tafsir pada ayat-ayat tersebut, maupun dengan metode lainnya yang dapat mengkaji tema dari ayat-ayat yang belum terklasifikasi.

Daftar Pustaka

Ibrahim, N. A., Kudus, A., Daud, I., & Bakar, M. A. (2008). Decision Tree for Competing Risks Survival. *nternational Journal of Biological and Medical Sciences*, 25-29.

Larose, D. T. (2005). *Discovering Knowladge in Data*. Canada: New Jersey.

Manning, C. (2010). *Text Classification and Naïve Bayes.* California: Stanford University.

Maria, R. (2013, April 02). *Pengeretian data mining, text mining dan web mining*. Diambil kembali dari E-COMMERCE: http://analisis-proses-bisnis-koperasi.blogspot.co.id/2013/04/pengertian-data-mining-text-mining-dan.html

Nusantara, B. (2015, Februari). *Metode Naive Bayes, sebuah penjelasan sederhana*. Diambil kembali dari LUKISANKEABADIAN: https://lukisankeabadian.blogspot.co.id/2015/02/metode-naive-bayes.html

Rozaq, A., Arifin, Z. A., & Purwitasari, D. (2012). Klasifikasi Dokumen Teks Berbahasa Arab Menggunakan Algoritma Naive Bayes. *ITS Surabaya*.