

Klasifikasi Kategori Hadits Menggunakan *Naive Bayes Classifier*

Emerald¹, Kemas Muslim Lhaksana, S.T., M.ISD²

^{1,2,3}School of Computing, Telkom University, Bandung, Indonesia

⁴Divisi Digital Service PT Telekomunikasi Indonesia

¹emraldkun@students.telkomuniversity.ac.id, ²kemasmuslim@telkomuniversity.ac.id

Abstrak [indonesia] Hadits merupakan salah satu panduan para umat Islam dalam melakukan aktivitas baik yang berkaitan dengan dunia maupun aktivitas yang berkaitan dengan akhirat. Sejauh ini hadits memiliki bermacam-macam kategori, dan kategori yang digunakan adalah tingkat keaslian hadits. Pada tugas akhir ini dilakukan penelitian untuk mengelompokkan pokok pembahasan dari kumpulan hadits menggunakan TF-IDF untuk menghitung bobot kata yang paling umum serta algoritma Naive Bayes untuk menghitung set dari probabilitas dengan cara perhitungan frekuensi dan kombinasi dari nilai didalam set data yang diberikan. Pengujian dilakukan sebanyak 2 kali, pengujian dengan data non pre-processing dan data setelah pre-processing. Dari hasil pengujian, nilai akurasi terbaik menggunakan data pre-processing yang memperoleh akurasi sebesar 82.27%.

Abstract [english] Hadith is a guide for Muslims to carry out both activities related to the world and activities related to the afterlife. So far the hadiths have various categories, and the categories used are the authenticity of the hadith. In this final project, a study was conducted to classify the subject matter of the hadith collection using TF-IDF to calculate the most common word weights and the Naive Bayes algorithm to calculate the set of probabilities by calculating the frequency and combination of values in the given data set. The test was carried out 2 times, testing with a non-preprocessing data and with a pre-processing data. From the test results, the best accuracy value uses the pre-processing data which has an accuracy of 82.27%

1. Pendahuluan

Latar Belakang

Hadits merupakan salah satu panduan para umat Islam dalam melakukan aktivitas baik yang berkaitan dengan dunia maupun aktivitas yang berkaitan dengan urusan akhirat. Hadits merupakan sumber hukum agama yang kedua setelah kitab suci Al-Qur'an. Istilah hadits pada awalnya berasal dari kata "Al-Hadits" dengan arti perkataan, percakapan, ataupun berbicara. Maka dapat disimpulkan bahwa hadits adalah setiap tulisan yang berasal dari perkataan atau percakapan Rasulullah SAW [1].

Hadits sejauh ini memiliki bermacam-macam kategori diantaranya, yaitu hadits berdasarkan keutuhan rantai sanad, jumlah penutur, dan tingkat keaslian hadits. Pada tugas akhir ini, kategori yang digunakan adalah tingkat keaslian hadits, secara lebih lengkap tingkat keaslian hadits terbagi atas hadits Shahih, Hasan, Dhaif, dan Maudlu' [1].

Pada tugas akhir ini dilakukan penelitian untuk mengelompokkan pokok pembahasan dari kumpulan hadits agar dapat digunakan oleh khalayak yang membutuhkan dengan menggunakan metode TF-IDF dan metode *Naive Bayes Classifier*, adapun kegunaan metode TF-IDF disini untuk menghitung bobot kata yang paling umum [2] dalam ayat di hadits, lalu metode *Naive Bayes Classifier* untuk perhitungan probabilitasnya.

Topik dan Batasannya

Berdasarkan latar belakang, maka berikut adalah rumusan masalah yang diangkat dalam penelitian tugas akhir ini :

1. Bagaimana klasifikasi kategori Hadits Menggunakan *Naive Bayes Classifier* ?
2. Bagaimana kinerja sistem yang dibangun dilihat dari akurasi *Success rate* nya ?
3. Bagaimana hasil performansi dari algoritma *Naive Bayes Classifier* dalam mengidentifikasi hadits ?

Adapun batasan masalah di dalam Tugas Akhir ini, yaitu pada percobaan ini menggunakan data yang didapatkan dari situs web *sunnah.com*, dan secara spesifik mengambil Hadits milik Imam Bukhari yang bernama Hadits Shahih Bukhari, juga dikenal dengan *al-Jami al-Musnad as-Sahih al-Mukhtasar min Umur Rasulullah SAW*

wa Sunanihi wa Ayyamihi, adalah salah satu koleksi yang yang terbaik karena Bukhari menggunakan kriteria yang sangat ketat dalam menyeleksi Hadits, dan disusun menjadi file *.CSV agar mempermudah proses percobaan.

Tujuan

Berdasarkan rumusan masalah yang ada, tujuan dari percobaan tugas akhir ini adalah mengklasifikasikan sebanyak mungkin/seluruh hadits yang ada dari berbagai sumber menjadi kategori-kategori tertentu menggunakan algoritma *Naive Bayes Classifier* yang akan memudahkan orang-orang dalam mencari hadits sesuai dengan kebutuhannya.

2. Studi Terkait

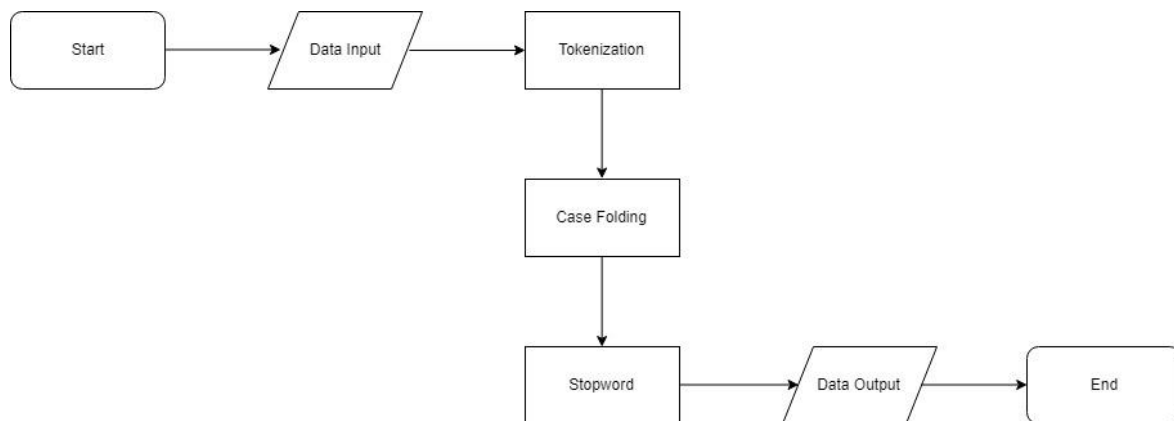
2.1 Hadits

Hadits merupakan salah satu panduan yang digunakan oleh umat islam dalam melaksanakan berbagai macam aktivitas baik yang berkaitan dengan urusan dunia maupun aktivitas yang berkaitan dengan urusan akhirat. Hadits merupakan sumber hukum agama islam kedua setelah kitab suci Al-Qur'an. Apabila ada perkara yang tidak dijelaskan dalam Al-Qur'an, maka umat islam akan menggunakan Hadits sebagai perantara keduanya. Istilah hadits pada dasarnya berasal dari bahasa Arab yaitu kata "Al-Hadits" yang artinya adalah perkataan, percakapan atau pun berbicara. Maka dapat dikatakan bahwa Hadits tersebut adalah setiap tulisan yang berasal dari perkataan atau pun percakapan Rasulullah Muhammad SAW [1].

Dalam sebuah hadits terdapat kategori tertentu yang membagi hadits tersebut, salah satunya adalah berdasarkan tingkat keaslian hadits nya, yang pertama yaitu hadits Shahih , hadits ini merupakan hadits yang sanadnya bersambung, paling diakui tingkat keasliannya dan paling banyak diterima oleh sekelompok ulama, lalu hadits Hasan , yang merupakan hadits dengan sanad bersambung, namun diriwayatkan oleh rawi yang tidak sempurna ingatannya, ketiga adalah hadits dhaif , hadits dengan sanad tidak bersambung ataupun diriwayatkan oleh rawi yang ingatannya tidak kuat, terakhir yaitu hadits maudlu', hadits ini merupakan hadits yang kurang lebih palsu atau hanya karangan dari manusia tanpa sumber yang pasti [1].

2.2 Pre-Processing

Tahap pre-processing adalah tahap dimana aplikasi melakukan seleksi data yang akan di proses pada setiap dokumen, pre-processing juga digunakan untuk mempermudah jalannya proses klasifikasi data. Berikut adalah flowchart serta penjelasan dari pre-processing yang dilakukan pada penelitian ini :



Gambar 1. Tahap Preprocessing

1. Tokenizing/Tokenization

Tahap ini adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya[3][4].

2. Case Folding

Case Folding dapat mengkonversi keseluruhan teks dalam dokumen menjadi sama rata (biasanya digunakan untuk menghuruf kecilkan atau lowercasing)[4].

3. Stopword

Stopword secara definisi adalah kumpulan kata-kata yang tidak berguna dan memiliki kekuatan diskriminasi rendah[5]. Tahap ini digunakan untuk mengecek apakah sebuah kata termasuk didalam

daftar yang harus dihentikan (Stoplist) seperti contoh “an, a, the, in” . Jika kata tersebut termasuk kedalam Stoplist, maka akan dihilangkan karena dianggap tidak ada kaitan dengan proses klasifikasi.

2.3 Machine Learning

Berdasarkan definisi yang dicetuskan oleh Arthur Samuel pada tahun 1959 machine learning adalah sebuah cabang ilmu yang memungkinkan komputer memiliki kemampuan untuk belajar tanpa harus di program lagi [6]. Machine Learning memiliki kaitan yang kuat dengan komputasi statistik dan juga berfokus pada prediksi dari penggunaan komputer. Dalam analisis data, Machine Learning digunakan untuk merancang sebuah algoritma serta model yang kompleks sesuai dengan prediksi. Hal ini dapat memungkinkan ilmuwan data, analisis, juga peneliti untuk mendapatkan hasil yang diinginkan melalui pembelajaran dari berkecenderungan dalam data dan hubungan yang historik.

Algoritma Naive Bayes adalah sebuah classifier probabilitas sederhana yang menghitung set dari probabilitas dengan cara perhitungan frekuensi dan kombinasi dari nilai didalam set data yang diberikan (suatu data dapat disebut kategori yang ditentukan apabila value nya lebih tinggi daripada kategori lainnya) . Algoritmanya menggunakan teorema Bayes dan menganggap semua atribut bersifat independen mengingat nilai variabel kelas. Independensi kondisional ini mengasumsikan bahwa jarang benar adanya pada penerapan di dunia nyata, maka karakteristiknya dapat dikatakan naif (*Naive*) namun algoritmanya cenderung memiliki kinerja yang baik dan belajar dengan cepat dari bermacam masalah klasifikasi yang lebih tinggi (*Supervised Learning*) [6]. Untuk proses klasifikasi kredibilitas informasi menggunakan teorema *naive bayes classifier* dengan teknik probabilitas *naive bayes classifier* untuk perhitungan data berbentuk kategori serta numerik, untuk sebuah dokumen h , dari semua kelas $x \in X$ classifier nya kembali menjadi kelas \hat{x} dimana memiliki probabilitas posterior terbesar yang diberikan dokumen. Berikut adalah persamaan teorema *naive bayes classifier* pada data berbentuk teks :

$$\hat{x} = \underset{x \in X}{\operatorname{argmax}} P(x|h) \quad (1)$$

Keterangan :
 \hat{x} : Estimasi dari kelas yang benar

Aturan *Bayes* ditampilkan pada rumus (2); yang memberikan cara untuk memecahkan segala probabilitas kondisional $P(x|h)$ menjadi tiga probabilitas :

$$P(h|x) = \frac{P(x|h) \cdot P(h)}{P(x)} \quad (2)$$

Keterangan :
 x : Data tes dengan kelas yang belum diidentifikasi
 h : Hipotesis data merupakan suatu kelas spesifik
 $P(h|x)$: Probabilitas h berdasar kondisi x (posteriori probability)
 $P(h)$: Probabilitas hipotesis h (prior probability)
 $P(x)$: Probabilitas x (*evidence*)

Dimana dalam percobaan ini h adalah kategori hadits yang digunakan, dan x adalah hadits itu sendiri, selanjutnya akan dilakukan substitusi dari rumus (2) kedalam rumus (1) untuk mendapatkan rumus (3) :

$$\hat{x} = \underset{x \in X}{\operatorname{argmax}} P(x|h) = \underset{x \in X}{\operatorname{argmax}} \frac{P(x|h) \cdot P(h)}{P(x)} \quad (3)$$

Keterangan :
 $P(x|h)$: Probabilitas x berdasar kondisi h (prior probability)
 $P(h)$: Probabilitas hipotesis h (prior probability)
 $P(x)$: Probabilitas x (likelihood)

2.4 Term Frequency – Inverse Document Frequency (TF-IDF)

Algoritma TF-IDF merepresentasikan hal penting dari sebuah kata. Term Frequency (TF) merepresentasikan berapa kali sebuah kata keluar didalam dokumen yang bersangkutan [7]. Semakin besar jumlah kemunculan suatu kata (TF tinggi) didalam dokumen, semakin besar juga bobotnya atau akan memberikan nilai kesesuaian yang semakin besar [7].

Pentingnya sebuah kata dalam sebuah dokumen dengan Term Frequency (TF) dapat dilihat di persamaan berikut :

$$TF(word) = \frac{Count(word)}{\sum_{i=0}^n Count(word_i)} \quad (3)$$

Keterangan :
 Count(word) : Jumlah kemunculan kata di dalam dokumen
 $\sum_{i=0}^n Count(word_i)$: Jumlah kemunculan semua kata di dalam dokumen

Inverse Document Frequency (IDF) adalah ukuran dari kemampuan kata untuk membedakan kategori. IDF dalam sebuah kata dapat diperoleh dari jumlah total dokumen yang terdapat kata tersebut dibagi oleh jumlah total dokumen setelah hasil bagi logaritmik [7]. Semakin sedikit jumlah dokumen yang mengandung kata yang dimaksud, maka nilai IDF semakin besar [7].

Berikut adalah persamaan dari Inverse Document Frequency (IDF) :

$$IDF(word) = \left(\frac{Count(docs)}{Count(word, docs)} + 0.01 \right) \quad (4)$$

Keterangan :
 Count(docs) : Jumlah dari dokumen
 Count(word,docs) : Jumlah dari dokumen yang memiliki kata tersebut

Inverse Document Frequency (IDF) yang sudah di hitung mengindikasikan bahwa jumlah dokumen yang memiliki kata tersebut lebih sedikit, $Count(word,docs)$ lebih kecil, maka mengindikasikan bahwa kata tersebut memiliki kelas diskriminatif yang sangat baik.

Term Weighting TF-IDF adalah penggabungan dari formula perhitungan $TF(word)$ dengan formula $IDF(word)$ []. Berikut adalah persamaannya :

$$Weight(word) = TF(word) * IDF(word) \quad (5)$$

Keterangan :
 $TF(word)$: Jumlah Term Frequency
 $IDF(word)$: Jumlah Inverse Document Frequency

2.5 Perhitungan Akurasi dengan Confusion Matrix

Akurasi merupakan sebuah sistem evaluasi dari model klasifikasi yang dibangun dan diharapkan dapat melakukan pengetestan dataset klasifikasi dengan benar, Confusion Matrix merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi. tetapi tidak dapat dipungkiri juga bahwa sistem kerja klasifikasi tidak dapat 100% benar sehingga sebuah sistem klasifikasi juga dapat diukur kinerjanya[9]

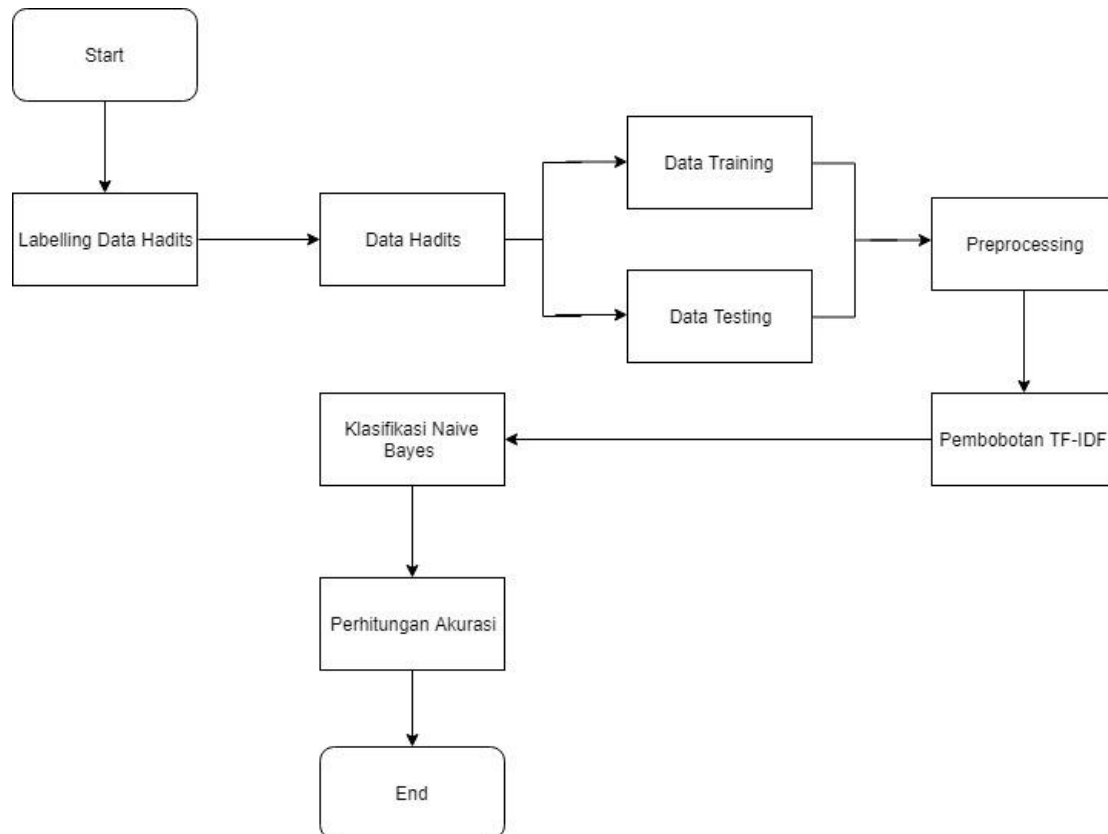
$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} * 100\% \quad (6)$$

Keterangan :
 TP (True Positive) : Jumlah data yang diprediksi benar oleh sistem
 TN (True Negative) : Jumlah data yang diprediksi salah oleh sistem

FP (False Negative) : Jumlah data negatif namun terprediksi benar oleh sistem
FN (False Positive) : Jumlah data positif namun terprediksi salah oleh sistem

3. Sistem yang Dibangun

Pada tugas akhir ini, dilakukan pre-processing data terlebih dahulu. Data akan terbagi menjadi dua yaitu data training dan data testing. Data yang digunakan merupakan data Hadits Shahih Al-Bukhari dari kumpulan Hadits-hadits di website sunnah.com. Data training nantinya akan dilatih dengan algoritma TF-IDF untuk mencari bobot dari setiap kalimat, dan selanjutnya data training yang sudah diolah akan di testing menggunakan algoritma Naive Bayes. Berikut ini adalah flowchart dari deskripsi sistem keseluruhan. Attribute class yang digunakan merupakan data yang sudah berada dalam waktu yang lama.



Gambar 2. Diagram pemrosesan data

3.1. Labelling

Labelling merupakan teknik pengambilan data secara manual dari suatu situs web. Data yang diambil adalah data Hadits Shahih Al-Bukhari. Data diunduh sebanyak 500 Hadits berbeda yang berasal dari sumber terpercaya sunnah.com. Selanjutnya data yang diambil akan di label dan dikategorikan sesuai dengan isinya.

3.2. Pre-Processing

Pre-processing pada penelitian ini yaitu melakukan beberapa Text pre-processing seperti tokenizing, stopwords, dan case-folding. Pada penelitian ini, dikembangkan beberapa atribut sebagai penunjang sistem kredibilitas. Berikut atribut yang dikembangkan berdasarkan penjelasan atribut pada tabel dibawah :

Tabel 1. Atribut pada sistem kredibilitas

Atribut	Penjelasan Atribut
<i>No. Hadits</i>	Nomor dari hadits sesuai dengan sumbernya
<i>Hadits</i>	Berasal darimana Hadits tersebut (Shahih apa)
<i>AyatHadits</i>	Ayat dari Hadits yang digunakan
<i>Kategori</i>	Pembeda antara satu hadits dengan yang lain berdasarkan isinya
<i>Index</i>	Jumlah Hadits per kategori

Jumlah atribut yang digunakan pada sistem sebanyak 5 atribut. Selanjutnya preprocess akan memasuki tahap TF-IDF (Term Frequency – Inverse Document Frequency) untuk mencari bobot dari kata yang digunakan dalam atribut.

Selanjutnya adalah tahap dalam Naive Bayes yang digunakan untuk menghitung probabilitas dari kata yang digunakan. Dan dilanjutkan dengan perhitungan akurasi dari keseluruhan data yang telah di konversi.

4. Evaluasi

4.1. Pengujian Sistem

Pada penelitian ini dilakukan beberapa pengujian data, dengan skenario pembagian data training dan data testing. Pengujian dilakukan sebanyak dua kali, yaitu pengujian bobot dari kata-kata menggunakan TF-IDF, dan selanjutnya menggunakan Naive Bayes Classifier untuk menentukan hasil akhir serta akurasi. Data tes tetap dalam jumlah yang sama baik dimulai dari pre-processing hingga tahap akhir atau tahap klasifikasi Naive Bayes.

4.2. Hasil Pengujian Data Pre-Processing

Pada pengujian data disini, dataset dibagi sesuai dengan skenario yang ada. Pengujian akan dilakukan sebanyak dua kali. Hasil pengujian beserta penjelasan dapat dilihat melalui tabel dibawah :

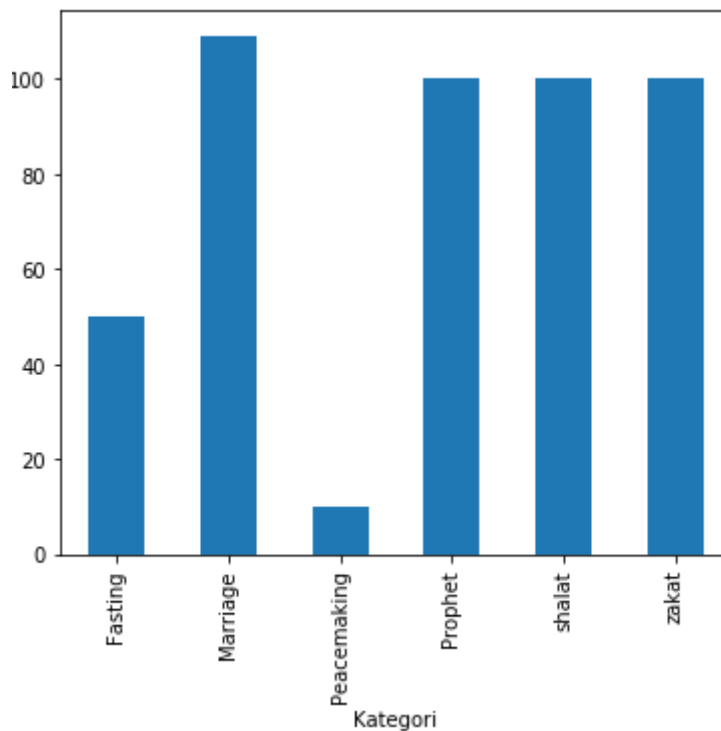
Tabel 2
Output Hadits Sebelum Pre-Processing

Ayat Hadits	Hadits	Kategori	Index
that he asked 'Aisha about the Statement of Al...	Shahih Al-Bukhari	Marriage	1.0
While I was with 'Abdullah, 'Uthman met him at...	Shahih Al-Bukhari	Marriage	2.0
We were with the Prophet (?) while we were you...	Shahih Al-Bukhari	Marriage	3.0
We presented ourselves along with Ibn 'Abbas a...	Shahih Al-Bukhari	Marriage	4.0
The Prophet (?) used to go round (have sexual...	Shahih Al-Bukhari	Marriage	5.0

Tabel 3
Output Hadits Setelah Pre-Processing

Ayat Hadits	Hadits	Kategori	Index
that he asked aisha about the statement of all...	shahih albukhari	Marriage	1.0
while i was with abdullah uthman met him at mi...	shahih albukhari	Marriage	2.0
we were with the prophet while we were young...	shahih albukhari	Marriage	3.0
we were presented along with ibn abbas at...	shahih albukhari	Marriage	4.0
the prophet used to go round have sexual rela...	shahih albukhari	Marriage	5.0

Berdasarkan kedua tabel diatas, Ayat Hadits beserta Hadits nya masih dalam posisi penuh dengan tanda baca dan huruf kapital sebelum pre-processing dimulai. Dan setelah pre-processing selesai , kedua atribut (Ayat Hadits, dan Hadits) telah bersih dari tanda baca dan huruf kapital. Selanjutnya adalah pengujian probabilitas dari setiap kategori yang dimiliki oleh data.



Gambar 3 Plot Probabilitas Kategori

Berdasarkan Plot diatas dapat ditunjukkan bahwa setiap kategori memiliki Probabilitas yang berbeda-beda tergantung dari banyaknya kategori tersebut ada didalam data yang digunakan.

4.3. Hasil Pengujian dengan menggunakan TF-IDF

Tabel 4
Output Pengujian TF-IDF

Kategori	Unigrams	Bigrams
Fasting	Fasting Fast	Fast day Month Ramadhan
Marriage	Marry Marriage	Said married Wedding banquet
Peackemaking	Following Peace	Said person Abdullah bin
Prophet	Jesus Abraham	Jews christians Son Mary
Shalat	Prayed Mosque	Used pray Prophet prayed
Zakat	Charity Zakat	Pbuh said Prophet pbuh

Berdasarkan hasil pengujian TF-IDF sesuai dengan tabel diatas, dapat ditunjukkan beberapa Kata (*Term*) dan Words (*Kalimat*) yang paling banyak frekuensinya keluar didalam sebuah hadits ditampilkan pada kolom bagian Unigrams dan Bigrams.

4.4. Hasil Pengujian dengan menggunakan Naive Bayes Classifier

Pada pengujian dengan menggunakan Naive Bayes Classifier, segala proses yang telah dilakukan pada metode dan tahap sebelumnya akan digabung dan akan dihitung akurasi akhir dari data penelitian yang digunakan. Berikut adalah hasil dari pengujiannya :

Tabel 5
Output pengujian NB Classifier sebelum Pre-Processing

Nilai akurasi dari skenario sebelum pre-processing (%)	Kategori	Probabilitas terprediksi setiap kategori
	Fasting	4.80108462e-03
	Marriage	4.16779789e-04
	Prophet	9.18610876e-03
	Shalat	5.44359914e-04
	Peacemaking	1.83672834e-04
	Zakat	9.84867994e-01
80.14%		

Berdasarkan hasil pengujian pada tabel 5, nilai akurasi yang diperoleh sebesar 80.14% , dengan probabilitas yang dimiliki kategori bermacam-macam. Selanjutnya akan dilakukan penelitian dengan data yang telah di Pre-Processing.

Tabel 6
Output pengujian NB Classifier setelah Pre-Processing

Nilai akurasi dari skenario setelah pre-processing (%)	Kategori	Probabilitas terprediksi setiap kategori
	Fasting	0.34938847
	Marriage	0.01621973
	Prophet	0.5759634
	Shalat	0.04263782
	Peacemaking	0.00590714
	Zakat	0.00988344
82.27%		

Berdasarkan hasil pengujian pada tabel 6, dari total 6 (enam) kategori beserta probabilitas yang berbeda dari masing – masing kategori tersebut, didapatkan akurasi sebesar 82.27% . Nilai akurasi mengalami kenaikan sebesar 2.13% dari pengujian tanpa pre-processing.

4.5. Analisis

Pada analisis penelitian ini, disebutkan bahwa hasil pengujian dengan melewati tahap pre-processing memberikan hasil yang lebih baik dibandingkan sebelum melewati tahap pre-processing. Hal ini disebabkan karena proses penentuan bobot dari sebuah Hadits terpengaruh dari adanya kata benda (punctuation) , huruf kapital yang tidak beraturan pada satu ayat hadits. Hal ini dapat menyebabkan nilai akurasi menjadi kecil.

5. Kesimpulan

Berdasarkan hasil pengujian diatas, bahwa Naive Bayes Classifier mampu mengukur akurasi dari kumpulan data Hadits yang telah dicari dan dibentuk sedemikian rupa. Dari hasil perbandingan akurasi antara non pre-process dan pre-process, nilai akurasi data diperoleh lebih besar dengan menggunakan pre-process dibandingkan dengan nilai akurasi non pre-process dari skenario data tes.

Untuk kedepannya (future works) nanti, sistem ini dapat dikembangkan lebih lanjut lagi, dengan cara-cara seperti berikut, menambahkan lagi jumlah data yang digunakan, menambahkan atribut pada data lagi, dan menggunakan metode yang familiar dengan penelitian diatas untuk mencari akurasinya.

Daftar Pustaka

- [1] *Pengertian dan Definisi*. (2014). Diambil kembali dari <https://pengertiandefinisi.com/pengertian-hadits-dan-jenis-jenis-hadits/>
- [2] Jie Chen, C. C. (2016). Optimized TF-IDF Algorithm with the Adaptive. 1 - 4.
- [3] *Python - Tokenization*. (2019). Diambil kembali dari Tutorials Point: https://www.tutorialspoint.com/python/python_tokenization.html
- [4] SB Kotsiantis, D. K. (2006). Data Preprocessing for supervised learning. *International Journal of Computer Science Volume 1*, 111-117.
- [5] *Removing stop words with NLTK in Python - GeeksForGeeks*. (2015). Diambil kembali dari GeeksForGeeks: <https://www.geeksforgeeks.org/removing-stop-words-nltk-python/>
- [6] Ristianto, W. (2016). *Warung Komputer*. Diambil kembali dari Machine Learning : Saat Komputer Belajar Memecahkan Masalah penting: <https://warungkomputer.com/2016/07/machine-learning-saat-komputer-memecahkan-masalah-penting/>
- [7] Hidayat, A. (2017). *Algoritma TF-IDF*. Diambil kembali dari <https://arfianhidayat.com/algoritma-tf-idf>
- [8] Aditya. (2016, November 12). *informatikalogi*. Diambil kembali dari <https://informatikalogi.com/term-weighting-tf-idf/>
- [9] Google Developers, 2018, "*Classification Accuracy*", Diambil dari <https://developers.google.com/machine-learning/crash-course/classification/accuracy/>
- [10] Hassan Saif, M. F. (2014). On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter. *oro.open.ac.uk*, 810-817.
- [11] Tina R. Patil, M. S. (2013). Performance Analysis of Naive Bayes and J48. *International Journal Of Computer Science And Applications*, 1-6.