KLASIFIKASI ANJURAN, LARANGAN, DAN INFORMASI PADA HADITS SHAHIH BUKHARI MENGGUNAKAN *NAÏVE BAYES CLASSIFIER*

CLASSIFICATION OF SUGGESTION, PROHIBITION AND INFORMATION OF SHAHIH BUKHARI HADITH USING NAÏVE BAYES CLASSIFIER

Syair Audi Liri Sacra[1], Said Al Faraby[2], Danang Triantoro M[3]

[1,2,3]Prodi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom

[1]syairaudilirisacra@gmail.com, [2]saidalfaraby@telkomuniversity.ac.id, [3]dto.lecture@gmail.com

Abstrak

Hadits adalah segala sesuatu yang dinisbatkan kepada Nabi Muhammad S.A.W baik berupa perkataan, perbuatan, taqrir (sikap diam setuju) dsb. Umumnya hadits di koleksi oleh beberapa imam besar, salah satunya koleksi hadits yang disusun oleh Imam Bukhari. Namun, saat ini belum ada penelitian maupun inovasi pengklasifikasian data berupa hadits berdasarkan anjuran, larangan dan yang hanya sekedar informasi. Dalam penelitian klasifikasi metode probablistik yang paling sering digunakan adalah Naïve Bayes Classifier karena memiliki keunggulan paling cepat dan sederhana. Naïve Bayes Classifier merupakan klasifikasi yang mengasumsikan keberadaan atribut suatu kelas tidak terkait atau tidak akan mempengaruhi atribut di kelas lain, atau dikenal dengan class conditional independence Namun permasalahan yang sering muncul pada penelitian terkait adalah tingginya dimensi data (banyaknya fitur / kata dalam satu opini). Hal ini tentu akan sangat mempengaruhi proses dari klasifikasi itu sendiri. Oleh karena itu, untuk mengurangi banyaknya fitur/kata pada sebuh data harus dilakukan beberapa tahapan seperti feature selection salah satu metodenya yaitu Chi-square. Penelitian ini menggunakan metode Naïve Bayes dengan menambahkan seleksi fitur pada penelitian ini menggunakan Chi-square..

Kata kunci : *Classification, Feature Selection, Naïve Bayes, Chi-Square, Hadits.*

Abstract

Hadith is everything attributed to the Prophet Muhammad in the form of words, deeds, Taqrir (silence agree) and so on. Generally hadith collections by some of the high priest, one a collection of hadith compiled by Imam Bukhari. However, there is currently no research and innovation classification of hadiths on the advice, prohibition and are merely information. In the classification probablistik method most frequently used is the Naïve Bayes classifier because it has the advantages of fast and simple. Naïve Bayes classifier is a classification that assumes the existence of the attributes of a class are not related or not will affect the attribute in another class, otherwise known as the class conditional independence However, the problems that often arise in related research is the high data dimensions (the number of features / word in one opinion). This will greatly affect the process of classification itself. Therefore, to reduce the high dimensional data has to be done several stages as one of the feature selection methods, namely Chi-square. This researvh using Naïve Bayes method by adding a feature selection in this study using Chi-square.

Keyword : Classification, Feature Selection, Naïve Bayes, Chi-Square, Hadith

1. Pendahuluan

Hadits adalah segala sesuatu yang dinisbatkan kepada Nabi Muhammad SAW baik berupa perkataan, perbuatan, taqrir (sikap diam setuju) dsb. Hadtis merupakan sumber hukum tersendiri bagi umat mulsim yang tidak dijelaskan dalam Al Qur'an[8] Setiap Hadist terdiri dari 2 bagian yaitu Sanad dan Matan. Sanad adalah untaian nama para penyampai Hadist yang menjamin keaslian dari Hadist itu. Matan sendiri adalah konten dari hadist itu. Setiap hadist diawali dengan Sanad[16]. Umumnya hadits di koleksi oleh beberapa imam besar, salah satunya koleksi hadits yang disusun oleh Imam Bukhari (nama lengkap: Abu Abdullah Muhammad bin Ismail bin Ibrahim bin al-Mughirah al-Ju'fi) yang hidup antara 194 hingga 256 hijriah.

Sebagai sumber hukum tersendiri, maka umat islam dianjurkan untuk mempelajari dan mengamalkannya. Hal tersebut dimudahkan dengan tersedianya cetakan buku kumpulan hadist hingga versi digital. Sejauh ini sumber – sumber tersebut hanya menyediakan hadist berdasarkan kitab pembahasannya namun hingga saat ini belum ada penelitian maupun inovasi pengklasifikasian hadits berdasarkan anjuran, larangan dan yang hanya sekedar informasi. Dengan adanya klasifikasi hadits akan memudahkan masyarakat yang akan mempelajari hadits berdasarkan ketegorinya.

Penelitian mengenai analisis klasifikasi pada kenyatannya telah banyak dilakukan. Naïve Bayes Classifier (NBC) ialah salah satu teknik pembelajaran mesin yang cukup sering digunakan untuk menangani hal tersebut. Seperti pada penelitian Twitter Sentiment Classification oleh Alec Go, dkk (2009) yang menggunakan Naïve Bayes dan dua metode lainny yaitu SVM dan MaxEnt menghasilkan nilai rata - rata akurasi terbaik dari NBC yaitu 81.375%, selanjutnya penelitian oleh Joko Samodra, dkk (2009) pada Klasifikasi Dokumen Teks Berbahasa Indonesia Menggunakan Naïve Bayes juga menunjukan metode NBC terbukti dapat memberikan hasil yang baik dalam melakukan klasifikasi dokumen teks berbahasa Indonesia, yang telihat dari akurasi yang terus meningkat hingga 87,63%.

Namun pada penerapannya Naïve Bayes Classifier merupakan klasifikasi yang mengasumsikan keberadaan atribut suatu kelas tidak terkait atau tidak akan mempengaruhi atribut di kelas lain, atau dikenal dengan class conditional independence. Sehingga akan menjadi celah untuk mengurangi keefektifan metode ini dan akibatnya meloloskan dokumen ke dalam kelas tertentu yang bukan kelas seharsunya.

Permasalahan yang sering muncul pada penelitian terkait adalah banyaknya fitur / kata dalam satu dokumen. Hal ini tentu mengganggu akan sangat mempengaruhi proses dari klasifikasi itu sendiri. Oleh karena itu, untuk mengurangi hal tersebut data harus dilakukan beberapa tahapan. Dalam hal tersebut biasanya metode yang dapat digunakan adalah feature selection. Feature selection adalah suatu proses untuk menyeleksi subset fitur dari sekumpulan fitur asli yang dapat mengurangi jumlah fitur, mempercepat proses algoritma data mining serta memperbaiki performansi data mining dengan cara menghilangkan fitur yang tidak relevan, redundan, noise. [12]. Selain itu dalam proses data mining, tahap preprocessing mengambil peranan penting terhadap kualitas data. Seperti stemming yang merupakan salah satu tahap dalam preprocessing yang bertujuan untuk mengubah suatu kata berimbuhan menjadi kata dasar. Pemotongan imbuhan tersebut akan berdampak pada perhitungan kemunculan kata yang selanjutnya dapat mempengaruhi perhitungan probabilitas suatu kalimat.

Oleh sebab itu pada penelitian ini akan dibahas mengenai hasil akurasi klasifikasi hadits Shahih Bukhari menggunakan metode klasifikasi Naïve Bayes dengan menggunakan feature selection Chi-Square serta bagaimana pengaruh tahap preprocessing stemming dalam proses klasifikasi data

2. Dasar Teori

2.1. Klasifikasi

Menurut KBBI klasifikasi merupakan penyusunan bersistem dalam kelompok atau golongan menurut kaidah atau standar yang berasal dari kata serapan bahasa Belanda, classificatie, yang sendirinya berasal dari bahasa Prancis classification. Istilah ini menunjuk kepada sebuah metode untuk menyusun data secara sistematis atau menurut beberapa aturan atau kaidah yang telah ditetapkan. Secara harafiah bisa pula dikatakan bahwa klasifikasi adalah pembagian sesuatu menurut kelas-kelas. Sedangkan menurut Ilmu pengetahuan, klasifikasi adalah proses pengelompokkan benda berdasarkan ciri-ciri persamaan dan perbedaan.

2.2. Data Preprocessing

Data yang digunakan dalam proses mining tidak selamanya dalam kondisi ideal untuk diproses. Terkadang pada data tersebut terdapat berbagai macam permasalahan yang dapat mengganggu hasil daripada data mining itu sendiri seperti diantaranya missing value, data redundant, outliers, ataupun format data yang tidak sesuai dengan system. Untuk mengatasi hal-hal tersebut maka diterapkanlah tahap Data Preprocessing [18].

Data preprocessing merupakan salah satu tahapan menghilangkan permasalahan-permasalahan yang dapat mengganggu hasil daripada proses data. Dalam kasus klasifikasi dokumen yang menggunakan data bertipe teks, terdapat beberapa macam proses yang dilakukan umumnya diantaranya case folding, filtering, stopword removal, stemming, tokenization, dan sebagainya yang akan dijabarkan pada bab tiga. [18]

2.3. Hadits

Hadits merupakan segala sesuatu yang dinisbatkan kepada Nabi Muhammad saw baik berupa perkataan, perbuatan, taqrir (sikap diam setuju) dsb. Pada dasarnya kedudukan Hadits dalam Islam, Al Sunnah atau al Hadits memiliki dua fungsi, yaitu sebagai berikut [8]:

- Mubayyin.

  Yaitu sebagai penjelas hal-hal yang disebutkan secara global dan umum dalam al Qur'an. Seperti; penjelasan tentang tatacara shalat, puasa, haji dsb. dan mengecualikan hal-hal yang umum dalam al Qur'an, seperti; Ahli Warits yang berhak menerima warits.

- Sumber Hukum tersendiri

  As Sunnah sebagai sumber hukum tersendiri dalam hal-hal yang tidak dibahas dalam al Qur'an baik secara global maupun terperinci, seperti; hukum haramnya menikahi dengan polygami ponakan dan bibinya, haramnya binatang yang bertaring, bercakar dsb.

2.4. Naïve Bayes Classifier

2.4.1. Teorema Bayes

Teorema Bayes merupakan sebuah algoritma klasifikasi statistik yang dapat memprediksi kelas suatu anggota probabilitas. Untuk klasifikasi Bayes sederhana yang lebih dikenal dengan nama Naïve Bayesian Classifier, dapat diasumsikan bahwa efek dari suatu nilai atribut sebuah kelas yang diberikan adalah bebas dari atribut-atribut lain. Asumsi ini disebut class conditional independence yang dibuat untuk memudahkan perhitungan-perhitungan, pengertian ini dianggap "naive", dalam bahasa lebih sederhana naïve itu mengasumsikan bahwa kemunculan suatu term kata dalam suatu kalimat tidak dipengaruhi kemungkinan kata - kata yang lain dalam kalimat padahal dalam kenyataanya bahwa kemungkinan kata dalam kalimat sangat dipengaruhi kemungkinan keberadaan kata-kata yang dalam kalimat. Secara matematis, teorema ini dapat diekspresikan sebagai berikut [4]:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (1)$$

Keterangan :

X :Data dengan class yang belum diketahui
H : Hipotesis data merupakan suatu class spesifik
P(H|X) :Probabilitas hipotesis H berdasar kondisi X (posteriori probabilitas).
P(H) : Probabilitas hipotesis H (prior probabilitas)
P(X|H) :Probabilitas X berdasarkan kondisi pada hipotesis H
P(X) : Probabilitas X

Untuk menjelaskan metode Naive Bayes, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut. Karena itu, metode Naive Bayes di atas disesuaikan sebagai berikut:

$$P(C|F1 \dots Fn) = \frac{P(C)P(F1 \dots Fn|C)}{P(F1 \dots Fn)})) \quad (2)$$

Di mana Variabel C merepresentasikan kelas, sementara variabel F1 ... Fn merepresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Maka rumus tersebut menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam kelas C (Posterior) adalah peluang munculnya kelas C (sebelum masuknya sampel tersebut, seringkali disebut prior), dikali dengan peluang kemunculan karakteristik-karakteristik sampel pada kelas C (disebut juga likelihood), dibagi dengan peluang kemunculan karakteristik-karakteristik sampel secara global (disebut juga evidence). Karena itu, rumus di atas dapat pula ditulis secara sederhana sebagai berikut:

$$Posterior = \frac{prior\ x\ likelihood}{evidence} \quad (3)$$

Nilai Evidence selalu tetap untuk setiap kelas pada satu sampel. Nilai dari posterior tersebut nantinya akan dibandingkan dengan nilai-nilai posterior kelas lainnya untuk menentukan ke kelas apa suatu sampel akan diklasifikasikan. Penjabaran lebih lanjut rumus Bayes tersebut dilakukan dengan menjabarkan $(C|F1, \dots, Fn)$ menggunakan aturan perkalian sebagai berikut:

$$\begin{aligned} P(C|F_{1,} \dots, F_n = \ & P(C)P(F_{1,} \dots, F_{n,}|C) \\ & = P(C)P(F_1|C)P(F_2, \dots, F_n|C, F_1) \\ & = P(C)P(F_1|C)P(F_2|C, F_1)P(F_3, \dots, F_n|C, F_1, F_1 \\ & = (C)P(F_1|C)P(F_2|C, F_1)P(F_3|C, F_1, F_2)P(F_4, \dots, F_n|C, F_1, F_{2,}F_{3)} \\ & = P(C)P(F_1|C)P(F_2|C, F_1)P(F_3|C, F_1, F_2) \dots P\left(F_n\middle|C, F_1, F_2, F_{3,} \dots F_{n-1}\right) \end{aligned} \quad (4)$$

Dapat dilihat bahwa hasil penjabaran tersebut menyebabkan semakin banyak dan semakin kompleksnya faktor - faktor syarat yang mempengaruhi nilai probabilitas, yang hampir mustahil untuk dianalisa satu persatu. Akibatnya, perhitungan tersebut menjadi sulit untuk dilakukan. Di sinilah digunakan asumsi independensi yang sangat tinggi (naif), bahwa masing-masing petunjuk (F1,F2...Fn) saling bebas (independen) satu sama lain. Dengan asumsi tersebut, maka berlaku suatu kesamaan sebagai berikut:

$$P\left(F_i\middle|F_j\right) = \frac{P(F_i \cap F_j)}{P(F_j)} = \frac{P(F_i)P(F_j)}{P(F_j)} = P(F_i) \quad (5)$$

Untuk i≠j, sehingga :

$$P\left(F_i\middle|C, F_j\right) = P(F_i|C) \quad (6)$$

2.4.2 Naïve Bayes Classifier

Naïve Bayes Classifier merupakan metode probabilistik yang mengasumsikan bahwa keberadaan atribut sebuah kelas tertentu tidak memiliki keterkaitan dengan atribut-atribut lain. Asumsi tersebut dikenal dengan class conditional independence yang dibuat untuk memudahkan perhitungan-perhitungan, sehingga perhitungan tersebut dianggap "naive", dalam implementasinya maksud naïve tersebut adalah bahwa kemunculan suatu term kata dalam suatu kalimat tidak dipengaruhi kemungkinan kata – kata yang lain dalam kalimat padahal dalam kenyataanya bahwa kemungkinan kata dalam kalimat sangat dipengaruhi kemungkinan keberadaan kata-kata yang dalam kalimat.

Salah satu kelebihan penggunaan Naive Bayes adalah metode ini hanya membutuhkan jumlah data latih yang sedikit untuk menentukan estimasi parameter yang diperlukan dalam proses klasifikasi serta menghasilkan akurasi yang tinggi.[15] Sehingga membuat klasifikasi ini prosesnya menjadi cepat dan sederhana.

Naïve Bayes menggunakan konsep dasar teorema Bayes, yaitu melakukan klasifikasi dengan melakukan nilai probabilitas, berikut adalah persamaan dari naïve bayes:

$$p(w_{kj}|c_i) = \frac{f(w_{kj},c_i)+1}{f(c_i)+|w|} \quad (7)$$

$f(w_{kj}, ci)$adalah nilai kemunculan kata w_kj pada kategori ci

$f(ci)$ adalah jumlah keseluruhan kata pada kategori ci

$|w|$ adalah jumlah keseluruhan fitur/kata yang digunakan

Dan

$$p(ci) = \frac{fd(ci)}{|D|} \quad (8)$$

$fd(ci)$adalah jumlah opini yang memiliki kategori ci

$|D|$adalah jumlah seluruh training opini

Dibentuk sebuah model probabilistic

$$p(w_{kj}|c_i) = \frac{f(w_{kj},c_i)+1}{f(c_i)+|w|} \quad (9)$$

Berdasarkan persamaan (9) di atas merupakan model dari teorema Naive Bayes yang selanjutnya akan digunakan dalam proses klasifikasi.

2.5. Feature Selection

Feature Selection adalah suatu proses untuk menyeleksi subset fitur dari sekumpulan fitur asli. Feature selection dapat mengurangi jumlah fitur, mempercepat proses algoritma data mining serta memperbaiki performansi data mining dengan cara menghilangkan fitur yang tidak relevan, redundan, noise, [1]. Dalam feature selection itu sendiri terdapat beberapa metode Chi-Square Chi-WSS, dan lain-lain.

2.6. Chi-square (x^2)

Chi-square(x^2) merupakan suatu perhitungan distribusi dalam statistika yang mengukur nilai ketergantungan antara term dan kategori [13]. Pada (Manning et al, 2009) menyatakan bahwa Chi-square(x^2) digunakan untuk menguji independensi antara dua kejadian yaitu kejadian kemunculan kata unuk dan kejadian kemunculan kelas. Jika nilai x^2< 10.83, maka tidak ada hubungan atau korelasi diantara dua variabel tersebut. Hal ini disesuaikan dengan persamaan dibawah ini :

$$x^2(W_k, C_i) = \frac{N_{(tr)}\,(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (10)$$

Dimana:

N (t_r) : jumlah dokumen dalam training set

A : jumlah dokumen dalam c_i yang mengandung w_k

B : jumlah dokumen yang bukan kategori c_i, tapi mengandung w_k

C : jumlah dokumen dalam c_i yang tidak mengandung w_k

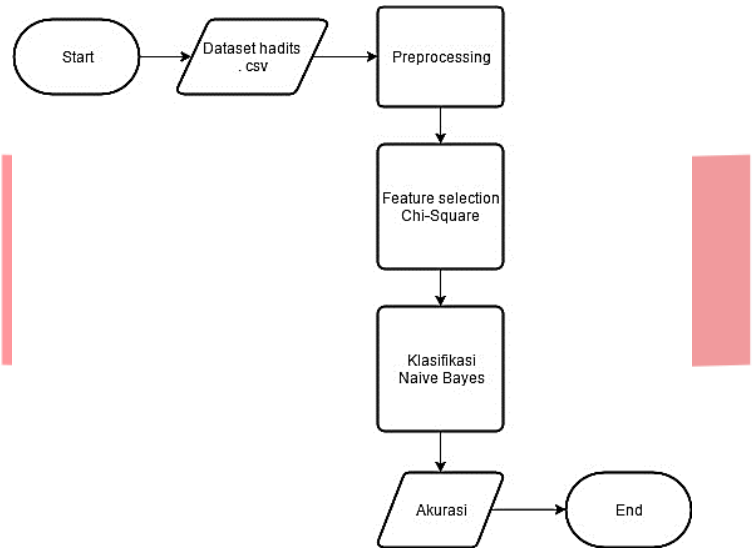
D : jumlah dokumen yang tidak dalam c_i dan tidak mengandung w_k

Ide dasar menggunakan Chi square untuk memilih fitur adalah untuk melihat korelasi fitur terhadap kategori. Perhitungan Chi square(x^2) dilakukan terhadap semua token pada setiap kategorinya. Token yang dipilih adalah token dengan nilai x^2 lebih besar atau sama dengan 10.83. Token dengan nilai yang tidak memenuhi syarat x^2 tidak digunakan sebagai fitur dalam proses learning. Hanya token yang memenuhi threshold yang ditentukan akan dipertahankan untuk digunakan dalam pemilihan fitur (Bramer, 2007).

3. Pembahasan

3.1 Gambaran Umum Sistem

Sistem yang dibangun pada penelitian tugas akhir ini adalah sebuah sistem yang dapat memberikan kalsifikasi berupa anjuran, larangan dan informasi pada tiga kelas dataset Hadits. Task yang dilakukan pada sistem ini terbagi menjadi tiga proses utama yaitu Preprocessing untuk mendapatkan data yang tidak mengandung noise yang dapat mengganggu proses berikutnya , feature selection menggunakan Chi-Square untuk menyeleksi dan mendapatkan fitur yang relevan saja serta klasifikasi menggunakan Naïve Bayes untuk mendapatkan probabilitas sehingga dapat diklasifikasikan berdasarkan nilai probabilitas kategori yang paling besar. Berikut adalah flowchart gambaran umum sistem yang akan dibangun pada penelitian ini:



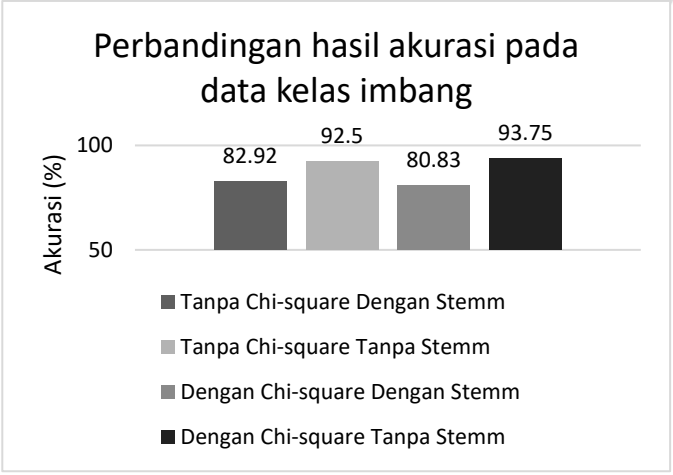
Gambar 1. Gambaran Umum Sistem

3.2 Analisis Hasil Observasi

3.2.1. Skenario Pengujian

Pada proses ini dilakukan pengujian dengan porsi 80:20 untuk data latih dan data uji. Pengujian klasifikasi anjuran, larangan dan informasi dilakukan sebanyak dua kali pada masing – masing jenis data. Yaitu data dengan porsi kelas yang seimbang, dan data dengan porsi kelas yang tidak seimbang. Untuk pengujian kelas seimbang digunakan data sebanyak 240 dengan pembagian 80 pada masing – masing kelas. Untuk pengujian kelas tidak seimbang digunakan data sebanyak 513 dengan pembagian anjuran 120, larangan 105, dan informasi 288. Pengujian dilakukan dengan data yang melewati proses stemming dan tanpa stemming. Pengujian dilakukan dengan klasifikasi naïve bayes tanpa feature selection chi-square dan dengan menggunakan feature selection chi-square dan Pengujian dilakukan dengan menerapkan K-Fold cross validation dengan nilai k = 10.

3.2.2. Pengujian data jumlah kelas imbang



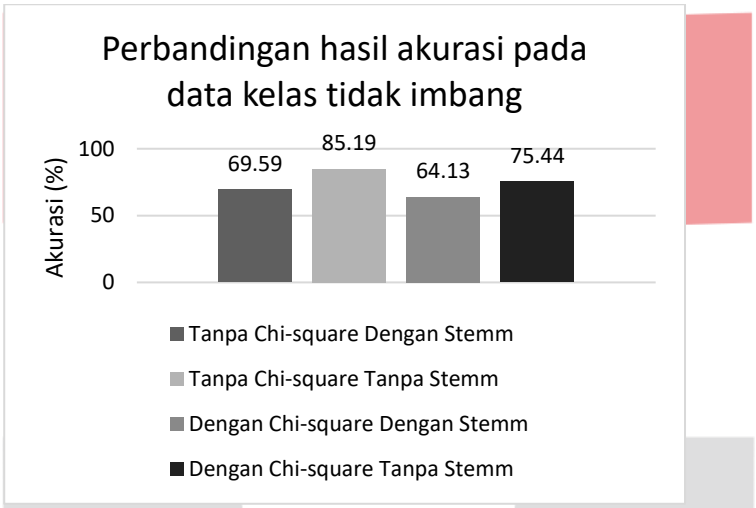
Gambar 2. Perbandingan hasil akurasi pada data kelas imbang

Berdasarkan gambar 1 dapat di ambil kesimpulan akurasi klasifikasi naïve bayes pada pengujian data imbang dengan menggunakan chi-square dan tanpa menggunakan stemming menunjukan hasil akurasi yang paling baik dibandingkan dengan pengujian lainnya dengan hasil sebesar 93.75%.

Namun dari pengujian yang dilakukan, pengujian klasifikasi naïve bayes tanpa menggunakan stemming menghasilkan akurasi yang lebih baik dibandingkan dengan menggunakan stemming, baik pada saat menerapkan teknik feature selection chi-square maupun tidak menggunakannya.

Hal ini dikarenakan pada proses stemming pemotongan imbuhan pada suatu kata mempengaruhi frekuensi kemunculan kata tersebut pada suatu kelas. Sehinggal hal ini akan mempengaruhi pula perhitungan pada klasifikasi naive bayes.

3.2.3. Pengujian data jumlah kelas tidak imbang

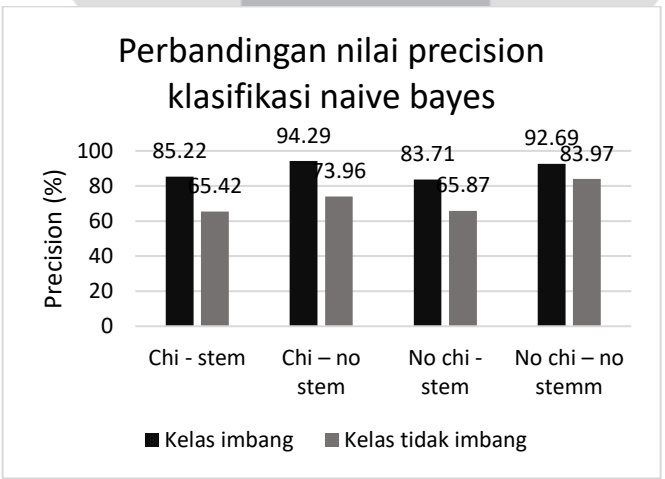


Gambar 3 Perbandingan hasil akurasi data kelas tidak imbang

Berdasarkan gambar 3 akurasi klasifikasi naïve bayes tanpa menggunakan chi-square untuk data kelas tidak imbang pada pengujian tanpa menggunakan stemming menghasilkan akurasi yang paling baik dengan nilai sebesar 85.19% dibandingkan dengan pengujian lainnya. Namun, dari pengujian yang dilakukan, pengujian klasifikasi naïve bayes tanpa menggunakan stemming menghasilkan akurasi yang lebih baik dibandingkan dengan menggunakan stemming, baik pada saat menerapkan teknik feature selection chi-square maupun tidak menggunakannya. Hal ini dikarenakan pada proses stemming pemotongan imbuhan suatu kata akan menyebabkan suatu kata unik dibaca hanya sebagai satu kata.

3.2.4. Evaluasi performansi precision, recall, dan f-measure

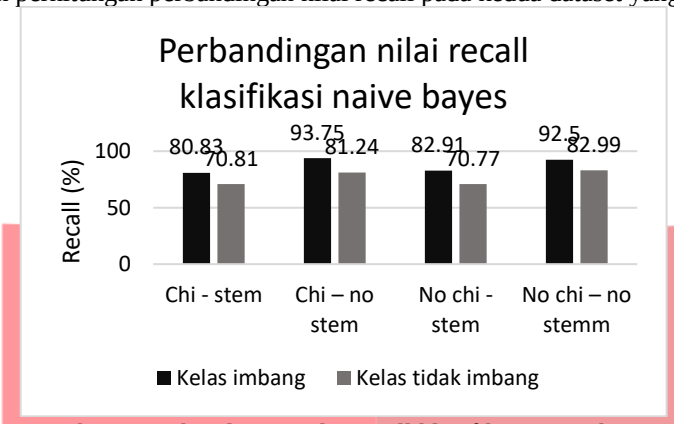
Selain melakukan perhitungan nilai akurasi, juga dilakukan perhitungan nilai precision, recall dan f-measure untuk mengukur perfomasi klasifikasi yang dilakukan oleh sistem yang dibangun. Berikut adalah analisis performasi pada masing – masing kelas dan pengujian.



Gambar 4. Perbandingan nilai precision klasifikasi naive bayes

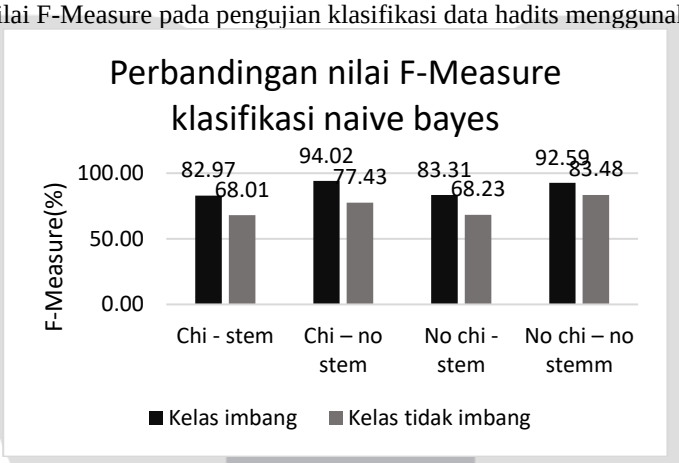
Berdasarkan Gambar 4 dapat dilihat bahwa nilai precision tertinggi adalah pengujian pada data dengan kelas imbang dengan menerapkan feature selection chi-square dan tanpa menggunakan proses stemming yaitu sebesar 94,29% . Hal

tersebut terjadi karena pada pengujian dengan jumlah kelas yang seimbang tidak ada suatu kelas mayoritas yang memberikan kecenderungan dalam proses klasifikasi sehingga mengurangi penyebab kesalahan dalam proses klasifikasi, serta dengan tidak dilakukannya proses stemming maka suatu kata yang menjadi kata unik dalam suatu kelas tidak terpotong imbuhannya. Seperti yang dijelaskan pada contoh tabel 4-6 dan 4-7. Perhitungan precision ini bertujuan untuk mengetahui seberapa banyak presentase prediksi klasifikasi yang di prediksi True dan terbukti True dibandingkan dengan jumlah keseluruhan prediksi bernilai True. Setelah dilakukan perhitungan perbandingan nilai precision, langkah selanjutnya adalah dilakukan perhitungan perbandingan nilai recall pada kedua dataset yang diujikan.



Gambar 5. Perbandingan nilai recall klasifikasi naive bayes

Berdasarkan Gambar 5 dapat dilihat bahwa nilai recall tertinggi adalah pada pengujian dengan menerapkan feature selection chi-square dan tanpa menggunakan proses stemming yaitu sebesar 93,75%. Perhitungan nilai recall dibutuhkan karena digunakan untuk mengetahui seberapa banyak presentase prediksi klasifikasi yang bernilai True dan terbukti True dibandingkan dengan keseluruhan data aktual yang bernilai True. Setelah dilakukannya perhitungan precision dan recall, untuk mengukur performansi kinerja sistem yang dibangun maka dilakukan perhitungan F-Measure (f1-score). Tabel 4-10 merupakan perbandingan nilai F-Measure pada pengujian klasifikasi data hadits menggunakan naïve bayes.



Gambar 6. Perbandingan nilai F-Measure klasifikasi naive bayes

Berdasarkan gambar 5 dapat dilihat bahwa terdapat kedua jenis dataset yang telah diuji menggunakan empat jenis pengujian serta telah dilakukan perhitungan precision dan recall pada halaman sebelumnya. Didapat hasil bahwa nilai F-Measure tertinggi untuk data dengan kelas imbang terdapat pada pengujian dengan menggunakan feature selection chi-square dan tanpa menggunakan stemming dengan nilai sebesar 94.02%, sedangkan untuk data dengan kelas tidak imbang nilai F-Measure tertinggi didapatkan pada pengujian tanpa menggunakan chi-square dan stemming dengan nilai sebesar 83.48%.

4. Kesimpulan

Berdasarkan hasil pengujian dan analisis yang telah dilakukan sebelumnya, maka dapat diambil kesimpulan sebagai berikut:

1. Klasifikasi Naïve Bayes saja tanpa feature selection akan menghasilkan akurasi yang lebih baik daripada Naïve Bayes dengan feature selection untuk dataset dengan kelas tidak imbang dan tanpa melalui proses stemming.sebesar 85.19%
2. Klasifikasi Naïve Bayes dengan feature selection akan menghasilkan akurasi yang lebih baik daripada Naïve Bayes tanpa feature selection untuk dataset dengan kelas imbang dan tanpa melalui proses stemming sebesar 93.75%.
3. Performansi terbaik pada penelitian ini didapatkan pada pengujian data dengan kelas imbang dengan menerapkan feature selection chi-square dan tanpa menggunakan proses stemming berdasarkan nilai F-Measure nya sebesar 94.62%
4. Porsi data pada tiap kelas mempengaruhi hasil akurasi klasifikasi Naïve Bayes

5. Saran

Adapun saran untuk untuk penelitian lebih lanjut mengenai klasifikasi Hadits maupun *feature selection* sebagai berikut:

1. Perlu dilakukan text preprocessing yang lebih baik agar data yang digunakan lebih berkualitas dan meminimalkan noise sekecil mungkin. Sehingga data yang digunakan memenuhi memenuhi kaidah dan ejaan yang benar.
2. Meningkatkan proses pelabelan data yang lebih baik untuk meningkatkan kualitas model klasifikasi.
3. Menerapkan teknik untuk menangani ketidakseimbangan data pada suatu kelas.

Daftar Pustaka:

[1] Permanasari, Wina, 2012, "Implementasi dan Analisis Orientation Detection pada Online Product Reviews dengan Naive Bayes dan Feature Selection", Institut Teknologi Telkom

[2] Hamzah, Amir. "Deteksi Bahasa untuk Dokumen Teks Berbahasa Indonesia." Seminar Nasional Informatika (SEMNASIF). Vol. 1. No. 1. 2015.

[3] Hidayatullah, Ahmad Fathan, and Azhari SN Azhari. "Analisis Sentimen dan Klasifikasi Kategori Terhadap Tokoh Publik pada Twitter." Seminar Nasional Informatika (SEMNASIF). Vol. 1. No. 1. 2015.

[4] O'Keefe, Koprinska. "Feature Selection and Weighting Methods in Sentiment Analysis." School of Information Technologies University of Sydney, 2006.

[5] Syahid, Bian, 2016 "Sentiment Classification pada Game Reviews", Universitas Telkom

[6] Ohana Bruno, 2009, Opinion Mining with the *SentiWordNet* Lexical Resource, Dublin Institute of Technology

[7] S.L, T., & W.H, A. (2011). Is Naive Bayes a Good Classifier for Document Classification ? *IJSEIA vol.5 No3*

[8] Zein, Muhammad Ma'shum. "Ulumul Hadits dan Musthalah Hadits." (2008).

[9] Liu, Bing. "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies 5.1 (2012): 1-167.

[10] Adnan, Husein, 2016, "Analisis Klasifikasi Sentimen dan Peringkasan pada Review Produk Menggunakan Metode Unsupervised", Universitas Telkom

[11] Triyuna, Mirza, "Perbandingan Klasifikasi Knn dan Naive Bayesian Serta Perbandingan Clustering Simple K-Means yang Menggunakan Distance Function Manhattan Distance dan Euclidian Distance Pada Dataset "Dresses_Attribute_Sales", Universitas Syiah Kuala.

[12] D. Tao and S. Wenqian, "An Improved Algorithm of Bayesian Text," Journal Of Software, 2011

[13] Abraham, Ranjit, 2009, "Effective Discreatization and Hybrid Feature Selection Using Naïve Bayesian Classifier for Medical Data Mining", Dr. MGR University, Chennai, India

[14] Tan, Ah-Hwee. "Text Mining: promises and challenges." South East Asia Regional Computer Confederation, Sigapore (1999).

[15] Pattekari, S. A., Parveen, A., 2012, Prediction System for Heart Disease Using Naive Bayes, International Journal of Advanced Computer and Mathematical Sciences, ISSN 2230-9624, Vol. 3, No 3, Hal 290-294.

[16] Naji Al-Kabi, M., Kanaan, G., Al-Shalabi, R., Al-Sinjilawi, S. I., & AlMustafa, R. S. (2005). Al-Hadith text classifier. Journal of Applied Sciences, 5, 584-587

[17] Tala, F. Z. (2003). A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. M.S. thesis. M.Sc. Thesis. Master of Logic Project. Institute for Logic, Language and Computation. Universiteti van Amsterdam The Netherlands.

[18] J. Han and M. Kamber, "Data Mining Concept and Technique".