Analisis Kelayakan Latar Belakang Proposal Penelitian Menggunakan Metode Naïve Bayes Classification

Feasibility Analysis of Background Research Proposals Using the Naïve Bayes Classification Method

Siska Febriani*[1], Muhamad Maksum Hidayat[2]
[1,2,3]Paska Sarjana Universitas Amikom Yogayakarta
E-mail: *[1]**siska17feb@gmail.com**, [2]**maksum.hidayat24@gmail.com**

Abstrak

Di akademisi atau kampus seringkali menjadi masalah bagi mahasiswa atau peneliti yang akan mengajukan proposal penelitian, di mana di latar belakang proposal yang diajukan seringkali tidak memenuhi atau tidak memenuhi unsur kelayakan latar belakang proposal, menurut Ristekdikti (2018)) latar belakang harus mengandung elemen latar belakang masalah, urgensi penelitian, dan tujuan spesifik penelitian. Ini menghasilkan sejumlah proposal penelitian yang tidak sesuai untuk persetujuan yang diakhiri dengan banyak penolakan atau revisi, hanya karena penulisan proposal penelitian yang tidak tepat. Dalam penelitian ini mengusulkan untuk menggunakan metode Naive Bayes sebagai metode klasifikasi teks latar belakang penelitian sehingga model yang sesuai diperoleh untuk menghasilkan tingkat akurasi terbaik untuk kelayakan latar belakang proposal penelitian.Dari hasil penelitian kelayakan latar belakang penelitian, Naïve Bayes Clasification (NBC) dapat digunakan untuk memproses klasifikasi data teks, terutama data latar belakang dari suatu penelitian. Dari beberapa dokumen yang traning data 3 layak, 2 tidak layak.

Kata Kunci*—Analisis, Kelayakan Latar Belakang, Naïve Bayes Classifier*

Abstract

In academia or campus it is often a problem for students or researchers who will submit a research proposal, where in the background of the proposal submitted often does not or does not fulfill the element of the feasibility of a proposal background, according to Ristekdikti (2018) the background must contain elements background problems, the urgency of the research, and the specific objectives of the study. This resulted in the number of inappropriate research proposals for approval ending with many rejections or revisions, only because of inappropriate research proposal writing.In this study proposes to use the Naive Bayes method as a method of classifying the background text of the study so that an appropriate model is obtained to produce the best level of accuracy for the feasibility of the background of the research proposal.From the results of the research background feasibility research, Naïve Bayes Clasification (NBC) can be used to process the classification of text data, especially background data from a study. Of the few documents that are data traning 3 feasible, 2 are not feasible.

Keywords*—Analysis, Background Feasibility, Naïve Bayes Classifier*

1. PENDAHULUAN

Saat ini di Internet ada banyak sekali sumber porlal jurnal yang menghasilkan banyak sekali jurnal akademis. Permintaan akan informasi oleh akademisi terus meningkat.Didalam dunia akademis atau kampus sering menjadi suatu permasalahan tersendiri bagi mahasiswa atau peneliti yang akan mengajukan sebuah proposal penelitian, dimana dalam latar belakang proposal yang diajukan sering belum atau tidak memenuhi unsur kelayakan sebuah latar belakang proposal, menurut Ristekdikti (2018) latar belakang harus memuat unsur latar belakang permasalahan, urgensi penelitian, dan tujuan khusus penelitian. Hal ini mengakibatkan banyaknya proposal penelitian yang tidak layak untuk disetujui berakhir dengan penolakan atau revisi yang banyak, hanya dikarenakan penulisan proposal penelitian yang tidak layak.

Berdasarkan hal tersebut diatas, penulis merencanakan untuk membuat sebuah aplikasi yang membantu mahasiswa atau peneliti untuk menguji kelayakan latar belakang proposal penelitian menggunakan metode *navie bayes*, sehingga meningkatkan kelayakan proposal yang akan diajukan dan meningkatkan kemungkinan penerimaan proposal penelitian.

Penambangan data adalah teknologi baru yang kuat untuk membantu perusahaan fokus pada informasi paling penting dalam data. Ini adalah proses menganalisis data dari berbagai perspektif dan meringkasnya menjadi informasi yang berguna. Saat ini perusahaan dan organisasi mengumpulkan data pada tingkat yang sangat tinggi dan dari tingkat yang sangat tinggi. beragam sumber seperti transaksi pelanggan, transaksi kartu kredit, penarikan tunai di bank.

Konsep dari aplikasi yang direncanakan penulis adalah kemampuan aplikasi untuk menganalisa dan menentukan dokumen proposal penelitian sudah memenuhi kriteria kelayakan sebuah latar belakang proposal atau belum, dengan mengklasifikasikan kalimat – kalimat yang menunjukan unsur-unsur penting dalam latar belakang. Dalam mengklasifikasikan kalimat dalam dokumen tersebut, aplikasi ini menggunakan metode algoritma *navie bayes*, dimana *Navie Bayes* merupakan salah satu metode yang banyak digunakan berdasarkan beberapa sifatnya yang sederhana, metode ini mengklasifikasikan data berdasarkan probabilitas.

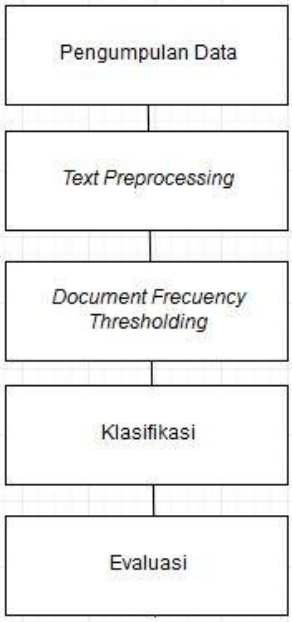
Penelitian tentang klasifikasi teks menggunakan *navie bayes* sudah banyak dilakukan oleh peneliti sebelumnya, seperti dilakukan oleh Wongso *et al*., (2017) melakukan penelitian untuk menemukan kombinasi mana pemilihan fitur dan klasifikasi yang dapat memberikan hasil terbaik untuk meningkatkan klasifikasi artikel koran di Bahasa Indonesia menggunakan kombinasi metode TFIDF dan *Multinomial Navie Bayes.* Somantri, (2017)melakukan penelitian klasifikasi teks untuk mengkategorikan cerpen menggunakan metode *navie bayes* dan membandingkan tingkat akurasinya dengan metode *Suppot Vector Mechine* (SVM). Kemudian Lichouri *et al*., (2018) melakukan penelitian tentang klasifikasi dialek arab dengan pendekatan tingkat kata dan tingkat kalimat dengan membandingan tiga metode klasifikasi teks yaitu *Linear Support Mechine* L-SVM , pendekatan tingkat kalimat *Bernoulli Naive Bayes BNB*, dan *Multinomial Naive Bayes* MNB.

Dari penelitian yang sudah dilakukan sebelumnya, perbedaan dengan penelitian ini adalah pada proses preprosesing data dan metode yang digunakan untuk klasifikasi teks latar belakang sebuah proposal penelitian. Berdasarkan dari kelebihan yang dimiliki maka pada penelitian ini mengusulkan untuk menggunakan metode *naive bayes* sebagai metode klasifikasi teks latar belakang penelitian sehingga didapatkan sebuah model yang tepat untuk menghasilkan tingkat akurasi yang terbaik untuk kelayakan latar belakang proposal penelitian.

2. METODE PENELITIAN

Pada bagian penelitian ini penulis menggunakan metodelogi penelitian *Action Research* dengan 4 tahapan yaitu tahap *planning, observation, action and reflection*. (Tanenhaus, M, 2005).

Tahapan-tahapan alur proses dari penelitian ini dijabarkan dalam gambar alur di bawah ini :



Gambar 1. Alur Penelitian

2.1 Metode Pengumpulan Data

Metode-metode yang dipakai untuk pengumpulan data dalam penelitian ini adalah mencari data dokumen latar belakang dari sumber yang terpercaya dan terindek untuk di lakukan pengolahan data pelatihan. Seperti dari Garuda, Sinta dan Scholarly. Dalam penelitin ini peneliti akan menggunakan 3 data latar belakang penelitian yang sudah terindek dan terakreditasi. Penelitin ini juga memebutuhkan contoh data latar belakang masalah yang akan di test kelayakannya.

Tahap yang kedua yaitu pembuatan program untuk melakukan proses *TextPrepocessing* dokumen. Menghasilkan data text yang di perlukan untuk tahap pelatihan.

Tahap yang selanjutnya yaitu *Document Frequency Thresholding.* menghitung*term* tiap data text yang diproses. Sehingga memperoleh data *term* tiap dokumen yang diproses dan menentukan batas atas dan bawah dari tiap *term.*

Tahap klasifikasi adalah pengelompokan data text pelatihan yang sesuai dan tidak dengan *naïve bayes classification. (Miner, 2012)*

Terakhir adalah evaluasi dari hasil proses data text yang sudah dilakukan. Jika belum memenuhi persyaratan yang sudah di tetapkan maka harus dilakukan penambahan data dokumen dan juga penentuan batas atas dan bawah tiap *term.*

2.2 NLP

Natural Language Processing (NLP) merupakan salah satu cabang ilmu Kecerdasan Buatan yang berfokus pada pengolahan bahasa natural. NLP dapat didefinisikan sebagai sebuah model yang memprogram interaksi antara pikiran dan bahasa (verbal dan nonverbal) sehingga dapat menghasilkan pikiran atau perilaku yang diharapkan.

2.3 Text Mining

Text mining atau data mining adalah proses dimana pengguna berinteraksi dengan koleksi dokumen menggunakan suatu sistem/ program. Data mining dan juga text mining adalah proses mencari informasi dengan pola tertentu yang berguna dari sumber agar hasil *output*nya dapat diolah lagidan diidentifikasi. (Imbar, 2014)

Dalam penelitian ini text mining bertujuan untuk mencari data text yang tidak terstrukturdiolah menjadi data yang terstruktur. Dalam pengolahannya nanti menggunakan algoritma data mining.

2.4 Text Preprocessing

Text Preprocessing adalah suatu proses mengubah bentuk data tekstual yang tadinya tidak terstruktur menjadi terstruktur. Dalam *text processing* ada beberapa proses yang terjadi(Sasmoyo, 2015). Proses-proses tersebut akan di jabarkan dalam Gambar 1 di bawah ini:

Gambar.2 Proses *Text Prepocessing*

2.5 Frequency Thresholding

Menurut beberapa penelitidokumen *frequency thresholding* adalah salah satu metode yang menghasilkan fitur terakurat dengan sistem kerja yang simple. Cara kerja dari perhitungan ini yaitu semakin banyak dokumen muncul (*term*tinggi) kemungkinan besar tidak memberikan kontribusi yang tinggi, maka dokumen itu bisa di hapus. Semakin sedikit dokumen *term* muncul maka semakin penting nilai data tersebut.

2.6 Naïve Bayes Clasification

Naïve Bayes Clasification adalah metode untuk mentukan probilitas prior bagi tiap kategori berdasarkan sampel dokumen. Dalam NBC terjadi dua proses klasifikasi yaitu, tahap pelatihan dan tahap klasifikasi. (Miner, 2016) Rumus mencari nilai maksimum yaitu :

$$V_{MAP} = \arg\max_{vj \in V} P(v_j \mid a_1, a_2, ..., a_n)$$

....[1] (Hamzah, 2012)

Bayes menyatakan tentang probabilitas bersyarat :

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

……[2] (Hamzah, 2012)

Nilai $P(v_j)$ ditentukan pada saat pelatihan, yang nilainya didekati dengan :

$$P(v_j) = \frac{|doc_J|}{|Contoh|}$$

…..[3] (Hamzah, 2012)

dimana|doc| adalah banyaknya dokumen yang memiliki kategori j dalam pelatihan, sedangkan |Contoh| banyaknya dokumen dalam contoh yang digunakan untuk pelatihan. Untuk nilai $P(w_{k|} v_j)$, yaitu probabilitas kata wkdalam kategori j ditentukan dengan:

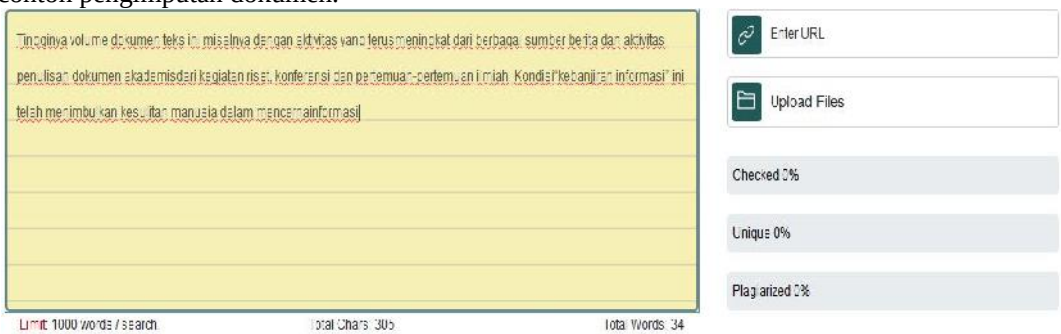
$$P(w_k \mid v_j) = \frac{n_k + 1}{n + |vocabulary|}$$ …[4](Hamzah, 2012)

Dimana n_k adalah frekuensi munculnya kata w_k dalam dokumen yang berkategori v_j, sedangkan nilai n adalah banyaknya seluruh kata dalam dokumen berkategori v_j, dan |*vocabulary*| adalah banyaknya kata dalam contoh pelatihan.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data dan Input Data

Proses pertama yaitu input data ke dalam aplikasi. User bisa memasukan lewat *copy paste*atau dengan mengupload file yang berbentuk word atau pdf. Di bawah ini adalah gambar contoh pengimputan dokumen.



Gambar 3. Contoh Dokumen Pelatihan

Dari data yang sudah diinputkan maka akan diproses menjadi data training dan data testing. Data yang digunakan pada awal training adalah ada 3 dokumen latar belakang yang sudah terindek dan terakreditasi. Setelah mendapat data training maka akan diproses kembali untuk mendapatkan parameter dan dilakukan proses input data latar bealkang ayang akan di testing. Dari hasil pencarian itu maka menghasilkan data testing.

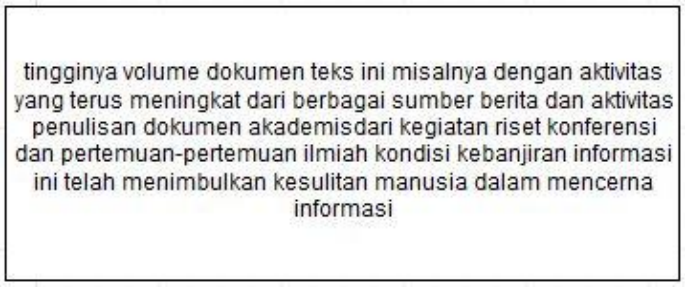
3.2 Text Prepocessing

Dalam *text preprocessing* terdapat 4 tahapan yang akan dilakukan. Pada penelitian kali ini contoh kutipan latar belakang yang akan di gunakan yaitu :

"Tingginya volume dokumen teks ini misalnya dengan aktivitas yang terusmeningkat dari berbagai sumber berita dan aktivitas penulisan dokumen akademisdari kegiatan riset, konferensi dan pertemuan-pertemuan ilmiah. Kondisi"kebanjiran informasi" ini telah menimbulkan kesulitan manusia dalam mencernainformasi" (Soematri, 2017).

1. *Case Floding*

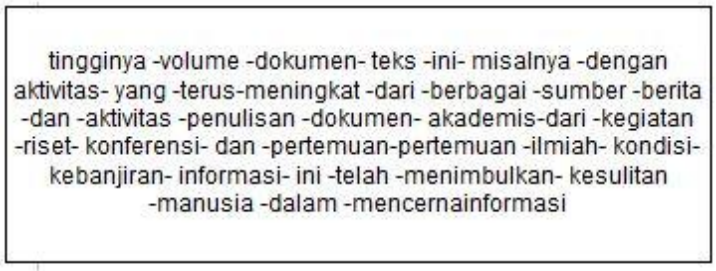
Setelah data diinput, maka data text yang akan di proses di jadikan huruf kecil semua. Tanda baca dalam text tersebut juga di hilangkan. Contoh *Case Floding* ada di gambar di bawah ini :



Gambar 4. Proses Case Floding

2. *Tokenization*

Tokenization adalah proses ke dua. Contohnya ada di gambar di bawah ini :



Gambar 5. Proses *Tokenization*

3. *Stemming*

Steaming yang di lakukan dalam penelitian ini ada di gambar berikut :

Gambar 6. Proses *Stemming*

4. *Filtering*

Ada beberapa kata yang di filter untuk mendapatkan kata-kata yang berguna dalam pengklasifikasian. Prosesnya ada di gambar berikut :

Gambar 7. Proses *Filtering*

3.3 *Document Frequency Thresholding*

Proses pertama setelah melakukan *preprocessing* yaitu menjadi pnilai kemunculan dari setiap kata yang ada. Di bawah ini dipaparkan tabel hasil dari tes 3 dokumen training.

Tabel 1. *Frequency Thresholding*

TF (TERM FREKUENSI)			
Q	D1	D2	D3
1	1	1	2
1	0	2	0
1	2	2	6
1	0	1	1
1	2	0	1
1	0	1	0
1	1	2	0
1	1	2	3
1	3	5	2

Pada tabel di atas adalah hasil TF-IDF dari 3 dokumen yang menjadi data training.

3.4 Klasifikasi

Setelah mengalami proses *TF-IDF* maka data training akan di cari nilai klasifikasinya. Berikut ini adalah tabel hasil nilai dari setiap dokumen training.

Tabel 2. Hasil Klasifikasi

Q*D1	Q*D2	Q*D3
0,000243662	0,000243662	0,000975
0	0,003846017	0
0,298995013	0,298995013	2,690955
0	0,000961504	0,000962
0	0	0
0	0,051822106	0
0	0	0
0,008211814	0,032847255	0,073906
0,672738779	1,868718832	0,298995
SUM(Q*D1)		
0,980189269	2,257434389	3,065793
0,527663307	0,788740228	0,917462

Dari hasil di atas diubah menajdi bentuk persen. Yaitu menjadi D1 52%, D2 78%, dan D3 91%. Yang menyatakan bahwa dokumen tersebut layak.

3.5 Evaluasi

Dalam proses evaluasi/pengujian ini peneliti mencoba menganalisis beberapa dokumen latar belakang yang akan di submit dalam jurnal terakreditasi. Berikut adalah tabel kemunculan data setelah di proses.

Table 3. kemunculan kata

kata kunci/Token	TF (TERM FREKUENSI)	
	D4	D5
banyak	0	6
kembang	0	0
tidak	0	2
jika	5	0
maka	0	0
besar	0	1
tujuan	0	0
hal	4	2
guna	0	0

Setelah daftar kata muncul dari proses klasifikasi, maka di lakukan pencarian TF-IDF dengan memasukkan beberapa dokumen baru yang belum di ketahui kelayakannya. Hasilnya di tabel berikut:

Tabel 4. Hasil TF-IDF

Q*D4	Q*D5
0	2,69095512
0	0
0	8,40889725
0	0
0	0
0	0
0	0
2,102224	2,10222431
0	0
2,102224	13,2020767
0,572278	2,63456103

Setelah mendapatkan nilai hasil Thresholdpeneliti mencoba menguji sistem dengan memasukkan beberapa dokumen. Hasilnya dadi tabel berikut :

Tabel 5. Hasil Threshold

Percobaan	akurasi	recall	F-Measure	Threshold
1	90%	78%	94%	
2	76%	45%	80%	
3	66%	55%	55.16%	0.0006
4	89%	34.78%	47.78%	0.0007
5	44%	89.11%	46.79%	0.0008
6	40%	69.76%	67.08%	0.0016

Dari tabel hasil threshold akan muncul hasil terhold perdokumen yang di coba. Setelah muncul nilai thresholdnya barulah data yang akan di ujicobakan di masukkan dalam sistem. Di bawah ini adalah hasil percobaan untuk uji kelayakan dari sebuah latar belakang yang sebelumnya belum diketahui.

Tabel 6. Pengujian

Threshold	Jumlah kata Per Kelas				
	banyak	kembang	tidak	jika	maka
0.0006	7776	3967	8844	7775	5578
0.0007	7633	5615	2345	3499	8798
0.0008	5834	7256	6443	9997	5147
0.0016	9921	3878	7854	3478	7456

Akhir dari penelitian ini adalah hasil layak atau tidak layak dari sebuah dokumen latar belakang. Di bawah ini tabel dari hasil dikumen di atas.

Tabel 7. Kelayakan

dokumen	hasil
1	tidak layak
2	tidak layak
3	layak
4	layak
5	layak
6	layak

Pada proses penginputan 300 dokumen di hasilkan 223 layak dan sisanya tidak layak. Dengan hasil ini dapat di pilih dokumen yang layak dan tidak layak dengan sistem yang kita teliti.

4. KESIMPULAN

Dari hasil penelitian kelayakan latar belakang penelitian, *Naïve Bayes Clasification*(NBC) bisa digunakan untuk memproses klasifikasi data teks khususnya data latar belakang dari sebuah penelitian. Dari ujicoba itu muncul nilai threshold nya yaitu 0.0006, 0.0007, 0.0008, 0.0016. dan setelah diujicobakan pada banyak dokumen yang akan di cari krlayakannya. Muncul kata banyak sebanyak 7776, 7633, 5834,9921.Dan setelah di input 300 dokumen latar belakang ada 223 layak dan sisanya tidak layak ada 77 dokumen.

5. SARAN

Saran untuk penelitian selanjutnya diharapkan menggunakan data training yang lebih banyak sehingga hasilnya akan lebih akurat. Selain itu data training sebaiknya di ambil dari banyak sumber yang berbeda sehingga akan lebih memperkaya pengetahuan sistem.

UCAPAN TERIMA KASIH

Peneliti mengucapkan terima kasih kepada semua pihak yang tidak bisa disebutkan satu persatu yang telah memberi dukungan moril maupun financial terhadap penelitian ini.

DAFTAR PUSTAKA

A. Hamzah, “Klasifikasi Teks Dengan Naive Bayes Classifier (NBC) untuk Pengelompokan Teks Berita dan Abstract Akademis,” dalam *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III*, Yogyakarta, 2012.

S. Sumpeno dan I. Destuardi, “Klasifikasi Emosi untuk Teks Bahasa Indonesia menggunakan Metode Naive Bayes,” *Seminar Nasional Pascasarjana*, 2009.

G. Miner, A. Fast, D. Delen, T. Hill, J. Elder dan B. Nisbet, Practical Text Mining and Statistical Analysis for Non-Structured Text Data Application, Oxford: Elsevier, 2012.

Lichouri, M. *et al.* (2018) ‘Word-Level vs Sentence-Level Language Identification: Application to Algerian and Arabic Dialects’, *Procedia Computer Science*. Elsevier B.V., 142(2017), pp. 246–253. doi: 10.1016/j.procs.2018.10.484.

Ristekdikti (2018) ‘Buku Panduan Pengusulan Program Penelitian Dan Pengabdian Kepada Masyarakat Melalui Simlitabmas Tahun 2018’, pp. 1–135.

Somantri, O. (2017) ‘Text Mining Untuk Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes (NB)’, *Jurnal Telematika*, 12(01). Available at: http://journal.ithb.ac.id/telematika/article/view/154.

Wongso, R. *et al.* (2017) ‘News Article Text Classification in Indonesian Language’, *Procedia Computer Science*. Elsevier B.V., 116, pp. 137–143. doi: 10.1016/j.procs.2017.10.039.

Rahma Amelia, *et al.* (2017) ‘Online News Classification Using Multinomial Naive Bayes. ITSMART: Jurnal Ilmiah Teknologi dan Informasi.,Vol. 6, No. 1, June 2017, ISSN:

2301-7201, E-ISSN: 2541-5689.

R. Imbar, Adelia, M. Ayub and A. Rehatta, "Implementasi Cosine Similarity dan Algoritma Smith-Waterman untuk Mendeteksi Kemiripan Teks," *Jurnal Informatika, 10(1),* 2014.

A. Sasmoyo, R. Saptono and Wiranto, "Penggunaan Jumlah Frekuensi Kata Terbanyak Sebagai Feature Set Pada Naive Bayes Classifier Untuk Mengklasifikasikan Dokumen Berbahasa Indonesia dan Inggris," *Seminar Nasional Ilmu Komputer,* 2015.