Jurnal Ilmiah Rekayasa dan Manajemen Sistem Informasi, Vol. 5, No. 2, Agustus 2019, Hal. 208-217
e-ISSN 2502-8995 p-ISSN 2460-8181

KLASIFIKASI PERMASALAHAN AGENSTOK MENGGUNAKAN ALGORITMA *NAIVE BAYES CLASSIFIER* PADA PT. HPAI-PEKANBARU

[1]Zarnelly, [2]Maya Andiany
Email: [1]zarnelly71@gmail.com, [2]maya.andiany@students.uin-suska.ac.id
[1,2]Program Studi Sistem Informasi, Fakultas Sains dan Teknologi,
Universitas Islam Negeri Sultan Syarif Kasim Riau

ABSTRAK

Kecendrungan seseorang untuk mengakses informasi khususnya permasalahan agenstok melalui dunia maya pun menjadi semakin tinggi. Informasi merupakan hal yang sangat penting dalam kehidupan masyarakat. Salah satu sumber infomasi adalah media sosial. Klasifikasi ini ditekankan untuk data permasalahan agenstok. Pada umumnya permasalahan yang disampaikan terdiri dari beberapa kategori seperti permasalahan mengenai kesehatan, konsultasi produk dan *marketing*. Namun dalam membagi permasalahan kedalan kategori-kategori tersebut untuk saat ini masih dilakukan secara manual.hal ini sangat merepotkan apabila permasalahan yang ingin di unggah berjumlah banyak. Oleh karena itu, perlu adanya sistem yang bisa mengklasifikasikan permasalahan secara otomatis. *Text mining* merupakan metode klasifikasi yang merupakan variasi dari data mining yang berusaha menemukan pola menarik dari sekumpulan data tekstual yang berjumlah banyak. Sedangkan algoritma *naive bayes classsifier* merupakan logartitma pendukung utuk melakukan klasifikasi. Kategori memiliki jumlah data permasalahan yang sama dan terdiri dari 400 data permasalahan; 360 data permasalahan digunakan untuk proses *training* dan 40 data permasalahan digunakan untuk proses *testing*. Pada penelitian ini metode yang digunakan yaitu *waterfall* dan pengujian *perfomance measure*, uji *black box*, dan uji sistem oleh pengguna. Adapun pengujian *perfomance measure* memperoleh nilai akurasi 97,5%, *precision* 97,6%, *recall* 97,5% dan *f-measure* 97,4%. Dari hasil-hasil tersebut dapat disimpulkan bahwa sistem yang menerapkan algoritma *naive bayes classifier* dapat digunakan untuk mengklasifikasikan permasalahan agenstok berbasis *web*, dengan menggunakan bahasa pemrograman *PHP* dan *Database Management System* (DBMS) menunjukkan bahwa klasifikasi permasalahan agenstok bisa terklasifikasi secara otomatis.

Kata Kunci: agenstok, akurasi, klasifikasi, *naïve bayes, text mining*

A. PENDAHULUAN

Teknologi informasi merupakan elemen vital dalam proses informasi yang dilakukan pada sebuah instansi atau perusahaan. Pentingnya peranan elemen tersebut telah mendorong terciptanya berbagai macam upaya untuk dapat memenuhi kebutuhan data dan informasi dengan melakukan *import* data yang cepat, efektif dan akurat. Salah satu contoh bidang yang membutuhkan teknologi informasi dalam kemajuan instansinya adalah bidang perdagangan dan jasa. Begitu pula PT. Herba Penawar Alwahida Indonesia, yang memiliki agen tersebar ke seluruh Indonesia yang memiliki peran penting sebagai konsultan bisnis dan sebagai agen terdepan dalam memasarkan produk yang mereka miliki. Dalam prakteknya, setiap orang bisa menjadi agen yang menjual produk.

Umumnya setiap agenstok memiliki beberapa masalah umum hingga spesifik sehingga untuk kasus yang sama harus diselesaikan secara berulang-ulang. Manajer akan memeriksa riwayat permasalahan yang ada di setiap masing-masing agenstok. Manajer merasa kesulitan untuk mengkategorikan permasalahan dari sistem konsultasi yang telah dibangun sebelumnya. Pada umumnya kategori permasalahan yang dikategorikan oleh konsultan terdiri dari beberapa kategori seperti mengenai kesehatan, konsultasi produk dan *marketing*. Namun dalam membagi permasalahan untuk kategori tersebut saat ini masih dilakukan secara manual. Salah satu sistem yang menangani konsultasi ini, awalnya menyediakan fitur kepada agenstok untuk mengisi keluhan-keluhan yang sedang dihadapinya.

Klasifikasi merupakan salah satu metode dalam data mining yang bertujuan untuk mendefenisikan kelas dari sebuah objek yang belum diketahui kelasnya. Pada klasifikasi terlebih dahulu akan dilakukan proses *training* dan *testing*. Klasifikasi dokumen bertujuan untuk mengelompokkan dokumen yang tidak terstruktur ke

dalam kelompok yang menggambarkan isi dari dokumen [1].

Dalam penelitian ini menggunakan metode berbasis statistik yaitu *naive bayes* yang memiliki kelebihan hanya memerlukan komputasi matematika yang tidak terlalu komplek sehingga sangat efesien dalam aplikasi praktis. Metode ini juga terbukti handal dengan tingkat akurasi cukup tinggi [2].

Pengklasifikasian permasalahan secara otomatis bisa dikategorikan sebagai *text mining*. Proses *text mining* dibagi menjadi tiga tahap, yaitu proses awal terhadap teks(*text preprocessing*), transformasi teks kedalam bentuk antara (*text transformation/feature generation)*, dan penemuan pola (*pattern discovery*) [3].

B. LANDASAN TEORI

B.1. Klasifikasi

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya kedalam kelas dari sejumlah kelas yang tersedia. Dalam klasifikasi ada dua pekerjaan utama yang dilakukan, yaitu pertama, pembangunan model sebagai *prototype* untuk disimpan sebagai memori dan kedua, penggunaan model tersebut untuk melakukan pengenalan, klasifikasi, prediksi pada suatu objek data lain agar diketahui kelas mana objek data tersebut dalam model yang sudah disimpan

Pada klasifikasi terdapat variabel target yang berupa nila kategorikal (nominal). Contoh dari klasifikasi adalah pendapatan masyarakat digolongkan kedalam tiga kelompok, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah. Algoritma yang biasa digunakan adalah *naive bayes*[4].

B.2. *Text Mining*

Text mining adalah suatu proses *knowledge-based* dimana pengguna berinteraksi dan bekerja dengan sekumpulan dokumen dengan menggunakan beberapa alat analisis.

Prosedur utama dalam metode ini terkait dengan menemukan kata-kata yang dapat mewakili isi dari dokumen untuk selanjutnya dilakukan analisis keterhubungan antar dokumen dengan menggunakan metode statistik tertentu seperti analisis kelompok, klasifikasi, dan asosiasi. Langkah-langkah yang dilakukan *text mining* [5].

1. *Text Preprocessing*

Text Preprocesing merupakan tahap tokenisasi yang merupakan proses pemecahan teks menjadi bentuk kata atau disebut sebagai token.

2. *Text Transformation/Feature Generation*

Pada tahap ini hasil yang diperoleh dari tahap *text preprocessing* akan melalui proses transformasi, dilakukan dengan mengurangi jumlah kata-kata yang ada dengan perhitungan kata-kata yang dianggap tidak penting (*stopword*)

3. *Stemming* Bahasa Indonesia

Dalam bahasa Indonesia, afiks/imbuhan terdiri dari sufiks(akhiran), infiks (sisipan), dan prefiks (awalan).

4. *Pattern Discovery*

Tahap penemuan polap*Pattern Discovery* adalah tahap terpenting dari seluruh proses *text mining*. Tahap ini berusaha menemukan pola atau pengetahuan dari keseluruhan teks.

B.3. *Term Frequency-Inverse Document Frequency* (TF-IDF)

Term Frequency-Inverse Document Frequency adalah statistik numerik yang mengungkapkan tingkat kepentingan kata sebuah dokumen dalam suatu koleksi. *Term Frequency-Inverse Document Frequency* sering digunakan sebagai faktor pembobotan dalam *information retrieval* dan *text mining*. Nilai TF-IDF meningkat secara proporsional berdasarkan berapa banyak kemunculan kata dokumen (*term frequency*), tetapi dinetralkan oleh frekuensi kata dalam *ccorpus (inverse document frequency)*.

Metode ini menggabungkan dua konsep untuk perhitungan bobot, yaitu frekuensi kemunculan kata didalam sebuah dokumen yang diberikan menunjukkan seberapa penting kata itu didalam dokumen tersebut. Rumus untuk TF-IDF [6]

B.4. *Confusion Matrix*

Confusion matrix adalah sebuah tabel yang menyatakan jumlah data uji yang salah diklasifikasikan. Berdasarkan jumlah keluaran kelasnya, sistem klasifikasi dibagi menjadi empat jenis yaitu, klasifikasi *binary*, *multi-class*, *malti-label* dan *hierarchial*. Pada klasifikasi *binary*, data masukan dikelompokkan kedalam salah satu dari dua kelas [7].

Tabel 1 *Confusion Matrix* untuk *multiclass*

		Kelas Prediksi		
		1	2	3
Kelas	1	TP	FN	TN

Sebenarnya	2	FP	TN	FP
	3	TP	FP	TN

Keterangan untuk tabel *confusion matrix*
True Possitive (TP_i), yaitu jumlah data positif yang terklasifikasi dengan benar oleh sistem untuk kelas ke-i.
True Negative (TN_i), yaitu jumlah data negatif yang terklasifikasi dengan benar oleh sistem untuk kelas ke-i.
False Negative (FN_i), yaitu jumlah data negatif namun terklasifikasi salah oleh sistem untuk kelas ke-i.
False Positive (FP_i), yaitu jumlah data positif namun terklasifikasi salah oleh sistem untuk kelas ke-i.

B.5. *Naïve Bayes Classifier* (NBC)

NBC adalah metode klasifikasi yang berdasarkan probabilitas dan *Teorema Bayesian* dengan asumsi bahwa setiap variable X bersifat bebas (*independence*). Dengan kata lain, NBC mengasumsikan bahwa keberadaan sebuah atribut (*variable*) tidak ada kaitannya dengan keberadaan atribut (*variable*) yang lain. Perhitungan perbandingan antara term pada *testing* dengan setiap kelas yang ada dengan menggunakan persamaan berikut [7]:

$$P(v_j) = \frac{n_c + mp}{n + m}$$

Keterangan:
n = jumlah term pada data latih dimana v = v_j
nc = jumlah term dimana v = v_j dan a = a_j
p = probabilitas setiap kelas dalam data latih
m = jumlah term pada data uji

Adapun untuk menentukan klasifikasi pada data uji digunakan persamaan:

$$V_{MAP} = argmax v_{j \in V} P(v_j) \prod P(a_i \vee v_j)$$

B.6. Phyton

Python adalah bahasa pemrograman yang fleksibel dan sederhana yang didefinisikan dalam dokumen-dokumennya sebagai berikut [8]:

1. Python adalah bahasa pemrograman tujuan umum yang sangat tingkat tinggi, dinamis, berorientasi objek, yang umum digunakan
2. bisa digunakan dalam aplikasi yang luas.
3. Bahasa ini dapat mendukung berbagai gaya pemrograman termasuk struktural dan berorientasi objek. Gaya lain juga bisa digunakan.
4. Python sangat fleksibel, karena kemampuannya untuk menggunakan komponen modular yang dirancang dalam bahasa pemrograman lainnya. Sebagai contoh, Anda dapat menulis sebuah program di C ++ dan mengimpornya ke python sebagai modul.

B.7. Basis data *(Database)*

Basis data dapat diartikan sebagai suatu pengorganisasian sekumpulan data yang saling terkait sehingga memudahkan aktifitas untuk memperoleh informasi. Basis data dimaksudkan untuk mengerti *problem* pada sistem yang memakai pendekatan berbasis berkas. Secara lebih lengkap, tujuan basis data adalah sebagai berikut:

1. Kecepatan dan kemudahan (*speed*)
2. Efesiensi ruang penyimpanan (*space*)
3. Keakuratan (*accuracy*)
4. Ketersediaan (*availability*)
5. Kelengkapan (*completness*)
6. Keamanan (*security*)
7. Kebersamaan pemakai (*sharability*)

B.8. *Software* Pendukung

Dalam merancang sistem berbasis *web* dibutuhkan beberapa aplikasi pendukung seperti:

1. MySQL

MySQL adalah sebuah *open source* sistem manajemen *database* SQL yang dikembangkan, didistribusikan oleh *Oracle Corporation.* Keistimewaan dari MySQL adalah dapat berjalan stabil di berbagai sistem operasi [9].

2. PHP

PHP adalah bahasa pemrograman *script side* yang didesain untuk mengembangkan web yang dapat ditanamkan ke dalam bahasa. Tujuan utama dari bahasa ini adalah agar pengembang web mampu menulis halaman web secara dinamis [10].

3. UML (*Unified Modelling Language*)

Notasi UML dibuat sebagai kolaborasi dari Grady Booch, DR. James Rumbough dan lainnya. UML menyediakan beberapa diagram yang menunjukkan berbagai aspek dalam sistem seperti *usecase diagram*, *activity diagram*, *sequence diagram,*dan *class diagram*. Tujuan perancangan UML adalah[11]:

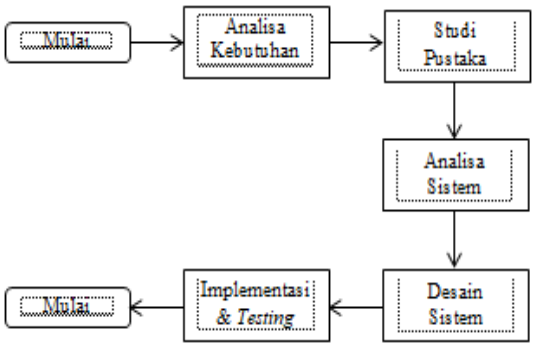
1. Menyediakan basis formal untuk pemahaman bahasa pemodelan

2. Mendorong pertumbuhan pasar kakas berorientasi objek.

C. METODOLOGI DAN DATA

C.1. Metodologi

Metodologi yang diterapkan dalam melakukan penelitian ditunjukan pada Gambar 1.



Gambar 1. Metodologi Penelitian

C.2. Kebutuhan Data

Sebelum melakukan proses klasfikasi, hal pertama yang dilakukan pada penelitian ini adalah mengumpulkan data permasalahan agenstok yang nantinya akan digunakan sebagai dataset. Data dikumpulkan secara manual melalui gruop *whatsapp*. Data permasalahan yang dikumpulkan berjumlah 1.200 data, waktu pengumpulan data permasalahan dengan rentang waktu bulan Januari sampai dengan Desember 2018 dan disimpan dengan format *xlsx. Selanjutnya menentukan kategori permasalahan yang akan dijadikan sebagai kelas pada proses klasifikasi menggunakan metode yang dibahas dalam penelitian ini. Adapun kategori dan ciri-ciri permasalahan adalah sebagai berikut:

1. Kesehatan

Beberapa permasalahan kesehatan ditandai dengan informasi yang ada dalam keluhan agenstok seperti berikut:

a. Cara untuk menggemukkan badan
b. Meningkatkan kecerdasan otak anak
c. Membantu pertumbuhan tinggi badan
d. Mengurangi kadar gula
e. Menjaga kesehatan mata

2. Konsultasi Produk

Beberapa contoh keluhan konsultasi produk ditandai dengan informasi yang ada dalam keluhan agenstok seperti berikut:

a. Madu yang bisa menguatkan daya hapal anak
b. Cara mengkonsumsi produk herbal
c. Khasiat produk herbal
d. Kegunaan promol2
e. Takaran pemberian obat herbal berdasarkan umur

3. Marketing

Beberapa contoh keluhan marketing ditandai dengan informasi yang ada dalam keluhan agenstok seperti berikut:

a. Syarat menjadi stokis
b. Daerah yang ada stokis dan pusat agensi
c. Cara menghitung bonus
d. Syarat pendaftaran menjadi agen
e. Cara menggunakan sistem AVO

C.3. Studi Pustaka

Tahapan ini merupakan tahapan dimana dilakukan informasi untuk lebih mengetahui tentang masalah dan teori-teori yang mendukung mengenai teori yang digunakan dalam penelitian ini. Studi pustaka dilakukan dengan membaca buku-buku, referensi, jurnal, dan penelitian sebelumnya yang membahas tentang klasifikasi, *text mining*, *dan* algoritma *naive bayes classifier*.

C.4. Analisa Sistem

Analisa sistem yang dilakukan adalah sebagai berikut:

1. *Pre-processing*

Tahapan yang dilakukan adalah:

a. *Case folding*, bertjuan untuk mengubah semua huruf kapital menjadi huruf kecil
b. *Tokenizing*, semua permasalahan agenstok yang telah dilakukan *case folding* akan dipisah menjadi token.
c. *Stemming*, konversi ke kata dasar.
d. *Filtering*, menyaring kata-kata yang tidak penting pada permasalahan agentok.
e. *Indexing*, membuat indeks kata dari permasalahan agenstok yang telah melalui proses tahapan sebelumnya.
f. Pembobotan kata, memberikan bobot pada kata yang telah indeks dengan menggunakan teknik pembobotan TF-IDF.

2. Analisa Sistem yang diusulkan

Penggunaan *text mining* untuk klasifikasi permasalahan menggunakan metode *naive bayes classifier* merupakan sebuah sistem yang dapat membantu mengklasifikasikan permasalahan menjadi lebih baik. Sistem ini dapat melakukan proses klasifikasi kategori masalah berdasarkan sistem informasi konsultasi agenstok berbasisi web yang telah dibangun pada penelitian sebelumnya. Kemudian sistem akan melakukan klasifikasi kategori permasalahan secara otomatis, agar data tersebut selanjutnya dapat disimpan kedalam *database*.

C.5. Desain Sistem

Dalam tahap ini ada empat perancangan yang akan dibuat, yaitu perancangan struktur menu dibuat menggunakan *tools* Ms. Visio 2007, Perancangan *interface* sistem yang akan dibuat menggunakan *tools* Mockup Balsamiq 3.5.7. Nota grafis yang menggunakan *Unified Modelling Language* (UML) meliputi *Usecase* diagram, *Activity* diagram, *Sequence* diagram, dan *class* diagram.

C.6. Implementasi

Implementasi merupakan tahap yang dilakukan setelah melakukan analisa dan perancangan. Data yang telah dianalisa sistem yang dirancang akan di implementasikan ke dalam bentuk tampilan dan koding. Pada penelitian ini, implementasi dilakukan dengan menggunakan laptop dengan spesifikasi perangkat keras yang dimiliki *processsor* Intel(R) Core(TM) i3 CPU M330@2.13 GHz, Kapasitas memori 4 GB, dan *Hardisk* 500 GB.

C.7.*Testing*

Setelah dilakukan implementasi, maka dilakukan pengujian terhadap sistem yang telah dibuat untuk mengetahui tingkat keberhasilan sistem yang telah dibangun menggunakan parameter pengujian yang ditentukan. Pada pengujian ini parameter yang digunakan adalah:

1. *Blackbox Testing*

Blackbox testing berkaitan dengan pengujian-pegujian yang dilakukan pada antarmuka perangkat lunak. Pengujian ini juga disebut pengujian fungsional karena pengujihanya melakukan pengujian pada perangkat lunak yang berkaitan dengan fungsionalitas dan bukan pada implementasi perangkat lunak. [14]

2. UAT (*User Acceptance Test*)

Pengujian yang dilakukan oleh pengguna dari sistem untuk memastikan fungsi-fungsi yang ada pada sistem tersebut telah berjalan dengan baik dan sesuai dengan kebutuhan pengguna. Hasil dari UAT adalah dokumen yang menunjukkan bukti pengujian dan diambil kesimpulan apakah *software* yang diuji dapat diterima [15]

3. Pengujian Akurasi, *Precision, recall dan F-measure.*

Perhitungan nilai akurasi didapat setelah mengetahui jumlah data uji yang benar diklasifikasikan. Perhitungan *precision* akan mengukur tingkat kepastian atau jumlah data *testing* yang diklasifikasikan dengan benar oleh model klasifikasi yang dibangun. *Recall* merupakan kebalikan dari *precision*. F-*measure* didapat dari perhitungan pembagian hasil dari perkalian *precison* dan *recall*, kemudian dikalikan dua [7].

D. ANALISIS DAN PEMBAHASAN

D.1. Pengumpulan Data

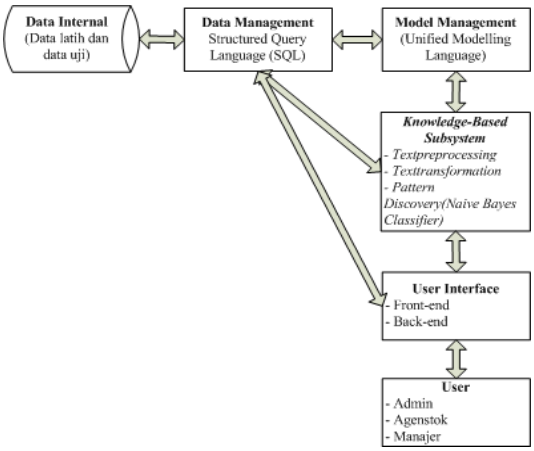
Tahap pengumpulan data menjadi tahap berikutnya pada penelitian ini yaitu tahap setelah dilakukan tahap perencanaan. Pengumpulan data didapaptkan dari *Whatsapp*. Total data yang dikumpulkan sebanyak 1.200 permasalahan data agenstok.

Tabel 2. Data Permasalahan Agenstok

No	Keluhan	Kelas
1	Kalau untuk menggemukkan badan apa ya?	Kesehatan
2	Saya kemarin sudah pernah daftar tapi belum ada dapat kartu anggota sampai sekarang, bagaimana?	Marketing
3	Aturan minum sari kurma untuk anak umur 3 tahun, bagaimana?	Konsultasi Produk
4	Assalamualaikum, untuk meningkatkan kecerdasan nutrisi otak anak apa	Kesehatan
5	Apa saja syarat untuk membangun pusat agensi	Konsultasi Produk
....	....	....
1200	Assalamualaikum, stokis daerah harapan raya dimananya?	Marketing

D.2. Arsitektur Sistem Klasifikasi Permasalahan

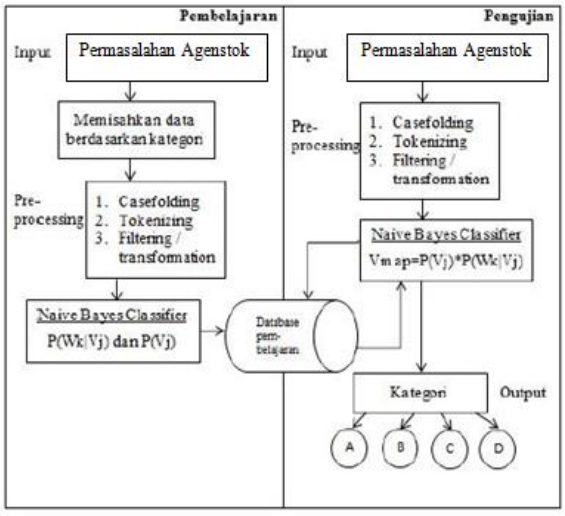
Berikut arsitektur sistem klasifikasi permasalahan agenstok terlihat pada Gambar 2.



Gambar 2. Arsitektur Sistem Klasifikasi Permasalahan

D.3. Gambaran Umum Sistem Klasifikasi

Sistem klasifikasi permasalahan agenstok ini merupakan suatu sistem yang mendukung pengelolaan data berbentuk teks secara otomatis dari hasil pengolahan data, informasi dan perancangan sistem. Dalam klasifikasi permasalahan ini, menggunakan metode *Naive Bayes Classifier* (NBC) dan proses yang paling penting dalam sistem ini yaitu penambangan sebuah teks pada suatu dokumen, sehingga dapat memberikan hasil yang sesuai dengan tujuan yang akan dicapai. Berikut gambaran umum sistem yang akan dibuat pada Gambar 3.



Gambar 3 Gambaran Umum Sistem Klasifikasi

Dapat dilihat pada Gambar 3 sistem yang akan dibangun terdapat dua proses yang berbeda yaitu proses pembelajaran(*training*) dan proses pengujian(*testing*).

1. Proses *training*, data permaslahan yang sudah dimasukkan akan dipisahkan berdasarkan kategori yang telah ditentukan. Kemudian data yang sudah dipisahkan masuk ketahap *preprocessing* (*casefolding*, *tokenizing*, dan *filtering*), setelah itu dihitung nilai probabibiltas kata dan probabilitas kategori pada setiap data yang dijadikan data *training*, kemudian disimpan ke dalam *database training* yang berisi kata-kata penting pada setiap kategori.
2. Sedangkan pada proses *testing* yang menjadi data masukkan yaitu data permaslaahan baru yang belum diketahui kategorinya. Pada tahap *preprocessing* yang dilakukan pada proses *preprocesing* didalam proses *training*, yang membedakannya yaitu pada saat perhitungan probabilitas setiap kata. Setelah melakukan tahap *preprocessing*, maka dokumen baru tersebut akan melalui proses persamaan kata. Kata-kata yang ada di data baru dengan kata-kata yang ada di proses *training*. Sehingga menghasilkan probabilitas pada setiap kategori yang ada.

D.4. Data *Training* dan Data *Testing*

Data latih (*data training*) dan data uji (*data testing*) diambil dari *dataset* awal dengan jumlah 1.200 permasalahan agenstok setelah melalui beberapa tahapan. Data tersebut dibagi untuk dilakukan pengujian menggunakan algoritma *Naïve Bayes Classifier*. Data dibagi menjadi data *training* dan data *testing* dengan perbandingan 90:10, dimana masing-masing kategori memliki data latih yang berjumlah 360 data permasalahan dan memiliki data uji di setiap kategori yang berjumlah 40 data permasalahan.

D.5. *Text Preprocessing*

Tahap *pre-processing* atau praproses data merupakan proses untuk mempersiapkan data mentah sebelum dilakukan proses lain. Pada umumnya, praproses data dilakukan dengan cara mengeliminasi data yang tidak sesuai atau mengubah data menjadi bentuk yang lebih mudah diproses oleh sistem.

Pada tahapan *filtering*, yaitu pembuangan kata-kata tidak penting dari hasil token. Selain itu juga dilakukan penghapusan tanda baca dan

stopword. *Stopword* diproses pada sebuah kalimat jika mengandung kata-kata yang sering keluar dan di anggap tidak penting seperti waktu, penghubung, dan lain sebagainya. Untuk itu perlu dilakukan penghapusan. Untuk melakukan proses penghapusan kata ini diperlukan sebuah data atau daftar kata yang diinginkan untuk dihapus. Adapun daftar kata yang digunakan adalah *stoplist* tala (Tala, 2003) dengan jumlah 758 *stopwords* [16]. Berikut adalah *stopword* yang digunakan yang dapat dilihat pada Tabel 3.

Tabel 3. Daftar *Stoplist* (Tala, 2003)

No	*Stopword*		*Stopword*		*Stopword*
1	Ada	11	Akhiri	21	Antaranya
2	Adalah	12	Akhirnya	22	Apa
3	Adanya	13	Aku	23	Apaan
4	Adapun	14	Akulah	24	Apabila
5	Agak	15	Amat	25	Apakah
6	Agaknya	16	Amatlah	26	Apalagi
7	Agar	17	Anda	27	Apatah
8	Akan	18	Andalah	28	Artinya
9	Akankah	19	Antar	...	...
10	Akhir	20	Antara	758	Yang

Pada tahapan *stemming*, yaitu pengubahan kata berimbuhan menjadi kata dasar dan pada tahapan ini menggunakan modul sastrawi pada python. Sedangkan pada tahapan *tokenizing*¸ setiap kata akan dipisahkan berdasarkan spasi yang ditemukan. Hasil *text preprocessing* dapat dilihat pada Tabel 4.

Tabel 4. Hasil *text preprocessing*

No	Keluhan	Kelas
1	untuk gemuk badan ya	Kesehatan
2	kemarin pernah daftar belum dapat kartu anggota sekarang	Marketing
3	atur minum sari kurma anak bagaimana	Konsultasi Produk
4	assalamualaikum tingkat cerdas nutrisi otak	Kesehatan
5	assalamualaikum saja syarat bangun pusat agensi	Marketing
6	testimoni promol12 tanam jagung	Konsultasi Produk
7	batuk masuk angin herbalnya	Kesehatan
8	assalamualaikum daerah kandang stok	Marketing
9	obat lebih sepesifik sakit satu obat	Konsultasi Produk
10	kapur sendi herbalnya	Kesehatan
...	....	...
1.200	ada stok harap raya	Marketing

D.6. *Term Frequency-Inverse Document Frequency* (TF-IDF)

Setiap keluhan dilakukan pembobotan dengan menggunakan persamaan. TF-IDF dihitung dengan ketentuan mengeliminasi *term* atau kata dengan frekuensi pada dokumen kurang dari 3% (Reza, 2017). Pada perhtiungan TF-IDF menggunakan bahasa pemrograman python.

Dengan melalui proses *text mining* yaitu proses *preprocessing* (*casefolding, tokenizing, dan filtering*), maka didapat jumah banyak kata sebanyak 5.215 dan jumlah frekuensi dari masing-masing kategori, jumlah frekuensi kategori kesehatan = 1.858, jumlah frekuensi kategori konsultasi produk = 2.030, jumlah frekuensi kategori marketing = 1.327.

D.7. TF-IDF Pada Data *Training*

Pembobotan TF-IDF pada 90% data *training* didapat sebanyak 5.215 kata. Kemudian hitung nilai probabilitas dengan menggunakan rumus:

$P(W_k|V_j)$ = (nk+1) / (Jumlah Frekuensi+Jumlah Kata)

Dimana:

$P(W_k|V_j)$ = Probabilitas bobot kata sesuai kategori

nk = Nilai kemunculan frekuensi kata (untuk mencari nilai kemunculan frekuensi didapat dari tabel *keyword*).

Prediksi : Apa obat untuk menjaga kesehatan mata

Setelah melewati hasil akhir yaitu stemming, kemudian dihitung nilai probabilitasnya kategori kesehatan, konsultasi prosuk dan *marketing* menggunakan rumus 4.14

Hasil probabilitas kategori kesehatan adalah:

P(obat| Kesehatan) = (121+1) / (1.858+5.215) = 0,0173

P(jaga| Kesehatan) = (3+1) / (1.858+5.215) = 0,0005

P(sehat| Kesehatan) = (2+1) / (1.858+5.215) = 0,0004

P(mata| Kesehatan) = (11+1) / (1.858+5.215) = 0,0016

Hasil probabilitas kategori konsultasi produk adalah:

P(obat| Konsultasi Produk) = (17+1) / (2.030+5.215) = 0,0024

P(jaga| Konsultasi Produk) = (2+1) / (2.030+5.215) = 0,0004

P(sehat| Konsultasi Produk) = (3+1) / (2.030+5.215) = 0,0005

P(mata| Konsultasi Produk) = (0+1) / (2.030+5.215) = 0,0001

Hasil probabilitas kategori *marketing* adalah:

P(obat| Marketing) = (0+1) / (1.327+5.215) = 0,0001

P(jaga| Marketing) = (0+1) / (1.327+5.215) = 0,0001

P(sehat| Marketing) = (0+1) / (1.327+5.215) = 0,0001

P(mata| Marketing) = (0+1) / (1.327+5.215) = 0,0001

Setelah mendapatkan nilai probabilitas kata pada setiap kategori, kemudian hitung probabilitas kategori dengan menggunakan rumus :

P(Vj) = Jml Dokumen setiap Kategori / Total Dokumen

Diketahui : Jumlah Dokumen Kesehatan = 360
: Jumlah Dokumen Konsultasi Produk = 360
: Jumlah Dokumen Marketing = 360

Jadi, probabilitas dari dokumen adalah :
P(Kesehatan) = 360/1200 = 0.3
P(Konsultasi Produk) = 360/1200 = 0.3
P(Marketing) = 360/1200 = 0.3

Sedangkan, pembobotan pada 10% data uji dapat dilihat pada Tabel 6 dengan *term* yang berbeda. Pembobotan ini digunakan untuk percobaan akurasi pada 10% data uji. Berikut ini perhitungan nilai probabilitas pada setiap kategori. Nilai yang dimasukkan berdasarkan data latih pada Tabel 5

1. Kategori Kesehatan

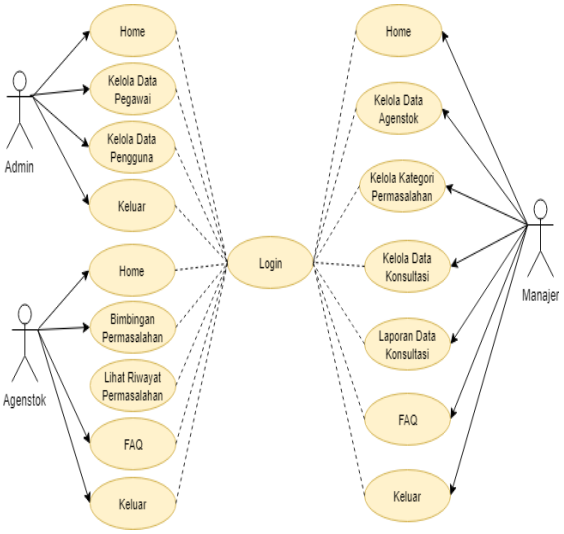
P(obat| Kesehatan) = 0,0173
P(jaga| Kesehatan) = 0,0005
P(sehat| Kesehatan) = 0,0004
P(mata| Kesehatan) = 0,0016
Jadi P(| Kesehatan) = 0, 0173 * 0, 0005 * 0, 0004 * 0,0016
= 0,00000000000554
Probabilitas = P(Kesehatan) * P(| Kesehatan)

= 0,3 * 0,00000000000554
= 0.00000000000116

Tabel 5. Probabilitas Data Uji

No.	Kesehatan	Konsultasi Produk	Marketing
1	5.34340^{-10}	2.29329^{-19}	4.38195^{-27}
2	0.001516	8.289394^{-06}	1.15003^{-27}
3	1.133337^{-05}	4.554612^{-08}	2.50007^{-38}
4	1.167052^{-06}	3.2793209^{-15}	3.00008^{-36}
...	...	...	...
120	8.333333^{-39}	1.4166667^{-37}	1.119698^{-05}

D.8. Deskripsi Fungsional Sistem

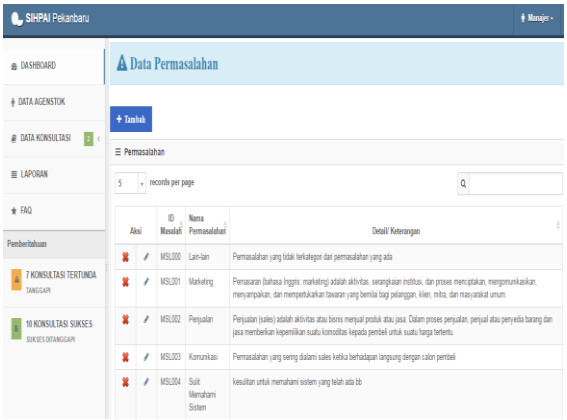


Gambar 4. *Usecase* Diagram

D.9. Implementasi dan Pengujian

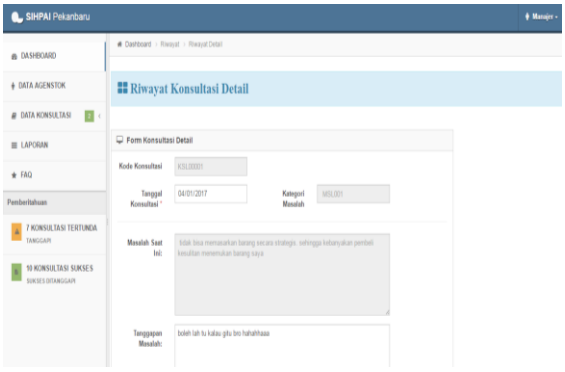
D.9.1. Implementasi

Tahap ini sistem siap dioperasikan oleh *user*. Pengguna sistem terdapat 3 aktor, yaitu admin, manajer, dan agenstok. Sistem klasifikasi pada aktor manajer terlihat pada Gambar 5.



Gambar 5. Kategori masalah

Tampilan *form* Klasifikasi Otomatis pada aktor manajer terlihat pada Gambar 6.



Gambar 6. Tampilan *Form* Klasifikasi Otomatis

D.9.2. Pengujian

Pengujian dilakukan untuk melihat hasil implementasi, apakah sistem berjalan sesuai tujuan atau masih terdapat kesalahan-kesalahan. Metode pengujian hasil klasifikasi dilakukan untuk mengetahui tingkat keakurasian sistem dengan menggunakan *confusion matrix* (akurasi) *precission*, *recall* dan *f-measure*, sedangkan metode pengujian sistem menggunakan *Blackbox Testing* dan *User Accaptance Test* (UAT). Bentuk pengujian ini untuk memastikan fungsi-fungsi yang ada pada sistem tersebut telah berjalan dengan baik dan sesuai dengan kebutuhan pengguna.

Pengujian Hasil Klasifikasi

Klasifikasi permasalahan pada PT. Herba Penawar Alwahida Indonesia melalui 2 tahapan utama, yaitu tahap pelatihan (*training*) yang masing-masing data dibagi dengan perbandingan 90% data pelatihan dan 30 % data pengujian seperti pada Tabel 6. Kemudian setelah proses klasifikasi selesai hasil klasifikasi diandingan dengan 4 pengujian yaitu akurasi, *precission*, *recall* dan *f-measure* untuk mengetahui data yang relavan sesuai kategori atau tidak.

Tabel 6. Jumlah Data *Training* dan Data *Testing*

No	Kategori	Data Pelatihan	Data Pengujian
1	Kesehatan	360	40
2	Konsultasi Produk	360	40
3	Marketing	360	40
Jumlah		1080	120
Total		1200	

Confusion Matrix

Confusion matrix adalah *tools* yang digunakan untuk evaluasi model klasifikasi untuk memperkirakan objek yang benar atau salah. Sebuah *matrix* dari prediksi yang akan dibandingkan dengan kelas asli dari inputan atau dengan kata lain berisi informasi nilai aktual dan prediksi pada klasifikasi.

Tabel 7. Klasifikasi dan Prediksi

Klasifikasi	Prediksi Kelas			
	Kesehatan	Konsultasi Produk	*Marketing*	
Kesehatan	40	0	0	40
KonsulProduk	3	37	0	40
Marketing	0	0	40	40
	43	37	40	

Adapun hasil pengujian klasifikasi tiga kategori, terlihat pada Tabel 8.

Tabel 8. Tabel Pengujian Klasifikasi

Evaluasi	Kesehatan	Konsul Produk	Marketing	Rata-rata
Akurasi	100	92,5	100	97,5
Precision	93,02	100	100	97,6
Recall	100	92,5	100	97,5
F-measure	96,3	96,1	100	97,4

Pada tabel akurasi diatas menampilkan hasil pengujian klasifikasi yang telah dilakukan, terdapat 4 pengujian dengan demikian dapat disimpulkan bahwa:

1. Nilai tertinggi terdapat pada pengujian *precision* dengan nilai 97,6%
2. Pengujian hasil klasifikasi pada akurasi dengan nilai 97,5 %
3. Pengujian hasil klasifikasi pada *recall* dengan nilai 97,5%
4. Pengujian hasil klasifikasi pada *f-measure* dengan nilai 97,4%

E. KESIMPULAN

Adapun setelah didapatkan hasil penelitian ini, kesimpulan yang didapatkan yaitu:

1. Sistem klasifikasi dibuat sesuai dengan algoritma yang digunakan yaitu algoritma *naive bayes classifier*. Dimana hasil pengujian klasikasi permasalahan yang dilakukan oleh sistem sudah sesuai dengan hasil klasifikasi secara manual.
2. Sistem mengklasifikasi permasalahan agenstok berdasarkan *query* konsultasi dengan menggunakan metode *naive bayes classifier* ini berhasil mengklasifikasikan permasalahan agenstok dalam kategori kesehatan, konsutasi produk dan marketing dengan tingkat rata-rata akurasi mencapai 97,5%.
3. Sistem menentukan klasifikasi permasalahan agenstok sesuai dengan hasil klasifikasi yang dilakukan secara manual dan sesuai dengan pengelompokkan permasalahan yang diinginkan oleh *user*

REFERENSI

[1] Februariyanti, H., & Zuliarso, E. (2012). *Klasifikasi Dokumen Berita Teks Bahasa Indonesia Menggunakan Ontologi*. Jurnal Teknologi Informasi. *Dinamik*, 17(1).

[2] Mahmudy, W, F., dan Widodo, A. W. (2015).. *Klasifikasi artikel berita secara otomatis menggunakan metode naive bayes classifier yang dimodifikasi*. *TEKNO*, 21(1).

[3] Sanjaya, S., dan Absar, E.A. (2015). *Pengelompokkan dokumen menggunakan winnowing fingerprint dengan metode k-Nearest neighbour*. *Jurnal Hail Penelitian Ilmu Komputer dan Teknologi Informasi*, 1(2), 50-56.

[4] Prasetyo, E. (2012). *Data mining konsep dan aplikasi menggunakan matlab*". Yogyakarta: Andi.

[5] Indriani, A. (2014). *Klasifikasi data forum dengan menggunakan metode naive bayes classifier*. *Jurnal Fakultas Hukum UII.*

[6] Nurjannah, M., Hamdani, & Astuti, I, F. (2016). Penerapan algoritma term frequency -inverse document frequency (tf-idf) untuk text mining". *Informatika Mulawarman*: *Jurnal Ilmiah Ilmu Komputer*, 8(3), 110-113.

[7] Purwanti, E. (2012). Klasifikasi dokumen temu kembali informasi dengan k-nearest neighbour". *Record and Library Journal*, 1(2).

[8] Nosrati, M. (2011). Phyton: An Appropriate Language For Real World Programming. W*orld Applied Programming*, 1(2), 110-117.

[9] Bartholomew, D. (2012). Mariadb vd. Mysql. *Dostopano*, 7(10), 2014.

[10] Ullman, L. (2011). *Php and mysql for dynamic web sites: Visual quickpro guide.* Peachpit Press.

[11] Sholiq, P .S. I. B. O. (2006). dengan uml. Yogyakarta: Graha Ilmu.

[12] Ratnawati. "*Pengembangan Aplikasi Profil Sekolah Berbasis Augmented Reality Sebagai Media Informasi Profil Sekolah di SMA Negeri 1 Wonogiri*". [Skripsi] Universitas Gunadarma. Yogyakarta. 2016

[13] Zarnelly, dan Adelia, D. (2015). Rancang bangun media pelayanan umum desk info info berbasisi web (Studi kasus: Pengadilan Tinggi Agama Pekanbaru). *Jurnal Informatika Rekayasa dan Manajemen Sistem Informasi*, 1(2), 55-59.

[16] Tala, F. Z.A. (2003). A *Study of Stemming Efects on Information Retrieval in Bahasa Indonesia, Institute for Logic, Language and Computation Van Amsterdam, The Netherlands.*