JSI : Jurnal Sistem Informasi *(E-Journal), VOL.12, NO.1, April 2020*
ISSN Print : 2085-1588
ISSN Online :2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index
email: jsifasilkom@unsri.ac.id

Klasifikasi Teks dengan *Naïve Bayes Classifier* (NBC) untuk Pengelompokan Keterangan Laporan dan Durasi *Recovery Time* Laporan Gangguan Listrik PT. PLN (Persero) WS2JB Area Palembang

Eka Afrianti[1], Fathoni[2], and Rahmat Izwan Heroza[3]
Sistem Informasi, Fakultas Ilmu Komputer Universitas Sriwijaya
Palembang, Indonesia
email : 09031181419008@students.ilkom.unsri.ac.id, fathoni@ilkom.unsri.ac.id,
rahmatheroza@unsri.ac.id

Abstrak

Di sebuah perusahaan, layanan sangat memengaruhi kepuasan pelanggan. Salah satu SLA (Service Level Agreement) yang PT. PLN (Persero) WS2JB Palembang Area memberikan tentang layanan, yang merupakan durasi waktu pemulihan harus kurang dari 180 menit. Namun dalam kenyataannya, durasi layanan sering melewati SLA. Untuk itu kita dapat mengklasifikasi laporan berdasarkan keterangan kedalam 3 kelompok. Berdasarkan penelitian yang telah dilakukan dengan menggunakan algoritma Naive Bayes Classifier dalam text mining diketahui bawah laporan yang teklasifikasi dibawah 180 menit ada 64,4% , 34,8% untuk laporan yang terklasifikasi kurang dari 1 hari dan 0,08% untuk laporan yang terklasifikasi lebih dari 1 hari.

***Kata Kunci* :** *Servis, PLN, Klasifikasi, Recovery Time, SLA, Text Mining, Naive Bayes Classifier.*

I. PENDAHULUAN

Listrik telah menjadi salah satu sumber energi utama dalam hidup kita untuk mendukung kegiatan sehari-hari. Mengacu pada Undang-undang Nomor 30 tahun 2006 tentang elektrifikasi adalah PT. PLN (Persero) sebagai satu-satunya BUMN untuk layanan listrik kami dan memiliki hak monopoli listrik di Indonesia. Jadi, tidak heran jika perusahaan ini memiliki banyak pelanggan seperti perumahan, bangunan, perkantoran dan industri.

Peningkatan kebutuhan listrik di Indonesia akan terus berlanjut, sebagai akibat dari peningkatan kualitas kesejahteraan dan perkembangan industri yang lebih cepat. Peningkatan listrik akan disertai dengan laporan kegagalan daya yang lebih banyak.

Pengaduan yang banyak ini harus diimbangi dengan mutu pelayanan seperti mengenai durasi pelayanan agar tidak mengecewakan pelanggan.

Di PT. PLN memiliki salah satu SLA durasi *recovery time* pelayanan pengaduan gangguan yang sebaiknya tidak melebihi 180 menit. Daftar data pengaduan gangguan tersebut jika diolah dengan *data mining*, maka akan menghasilkan pola data yang berguna untuk mengevaluasi pelayanan yang terjadi selama ini khususnya untuk durasi *recovery time*, seperti untuk memprediksi durasi *recovery time* rata-rata posko.

II. TINJAUAN PUSTAKA

A. Data Mining

Data Mining merupakan proses dalam menemukan pola atau informasi menarik dari sejumlah data yang besar, dimana data dapat disimpan dalam *database, data warehouse* atau dapat disimpan di tempat penyimpanan informasi lainnya dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika [3]. Secara besar data mining

dikelompokkan menjadi beberapa kategori yaitu [4]:

- Deskripsi
- Estimasi
- Prediksi
- Klasifikasi
- Pengklusteran
- Asosiasi

B. Text Mining

Text Mining adalah proses ekstraksi pola (informasi dan pengetahuan yang berguna) dari sejumlah besar data yang tak terstruktur. *Text mining* mempunyai tujuan dan proses yang sama seperti *data mining*, namun memiliki masukkan yang berbeda. Masukkan untuk *text mining* adalah data tekstual (*natural language*) yang tidak atau kurang terstruktur seperti dokumen *word*, pdf, kutipan teks, dll, sedangkan untuk *data mining* masukkannya berupa data yang terstruktur [2]. *Text mining* menggunakan penerapan *data mining* untuk mengubah data tidak terstruktur menjadi data terstruktur melalui tahap- tahap sebagai berikut : [2]

1. *Text Preprocessing* yaitu pemecahan karakter ke dalam kata-kata.
2. *Feature Selection* yaitu seleksi *feature* untuk mengurangi dimensi dari suatu kumpulan teks.
3. *Text Mining / Pattern Discovery* yaitu dapat berupa *unsurpervised learning (clustering)* atau *supervised learning (classification)*.
4. *Interpretation / Evaluation* yaitu pengukuran efektifitas

C. Naive Bayes Classifier (NBC)

Pengklasifikasi bayes merupakan salah satu pengklasifikasi statistik, dimana pengklasifikasi ini dapat memprediksi probabilitas keanggotaan kelas suatu data yang akan masuk ke dalam kelas tertentu, sesuai dengan perhitungan probabilitas. Pengklasifikasi Bayes didasari oleh teorema bayes yang ditemukan oleh Thomas Bayes. Pada abad ke-18. Dalam studi pembandingan algoritma klasifikasi telah ditemukan simple bayesian atau yang biasa dikenal dengan *Naïve Bayes classifier*. *Naïve Bayes classifier* menunjukkan akurasi dan kecepatan yang tinggi bila diterapkan pada database yang besar [2]. Metode ini sering digunakan dalam menyelesaikan masalah dalam bidang *machine learning* karena metode ini dikenal memiliki tingkat akurasi yang tinggi dengan perhitungan sederhana [1]. Teorema bayes merupakan dasar aturan dari *naive bayes classifier* berikut teorema bayes akan disajikan pada persamaan (1).

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

Dimana X merupakan data hasil pengujian dari suatu set data yang telah ditentukan masuk ke ke dalam kelas tertentu. H merupakan suatu hipotesis yang akan menentukan X masuk ke dalam kelas C. P(H|X) merupakan peluang atau probabilitas X yang merupakan data atau bukti

yang diperoleh pada saat observasi masuk ke dalam kelas C, dengan kata lain mencari probabilitas X dimiliki oleh kelas C. P(H|X) merupakan probabilitas posterior, H dikondisikan pada X. Sebaliknya P(H) merupakan probabilitas prior, atau probabilitas sebelumnya. Kemudian P(X|H) merupakan probabilitas posterior dimana X dikondisikan pada H. Sedangkan P(X) merupakan probabilitas sebelumnya dari X [2].

Dengan aturan Bayes maka penelitian ini akan mengimplementasikan aturan bayes dengan studi kasus tertentu oleh karena itu aturan bayes dapat dinyatakan :

$$P(C_j|X) = \frac{P(X|c_j)P(C_J)}{P(X)} \quad (2)$$

Dimana c jadalah kategori teks yang akan diklasifikasikan, dan p(c¿¿ j)¿ merupakan probabilitas prior dari kategori teks cj. Sedangkan d merupakan dokumen teks yang direpresentasikan sebagai himpunan kata (W1,W2, … Wn), dimana W1 adalah kata pertama, W2 adalah kata kedua dan seterusnya. Pada saat proses pengklasifikasian dokumen teks, maka pendekatan Bayes akan memilih kategori yang memiliki probabilitas paling tinggi (CMAP) yaitu :

$$C_{MAP} = argmax \frac{p(cj)\,p(X \vee c_j)}{P(X)} \quad (3)$$

Nilai $p(X)$ dapat diabaikan karena nilainya adalah konstan untuk semua *cj*, sehingga persamaan (3) dapat dituliskan :

$$C_{MAX} = argmax\, p(c_j)p(X \vee c_j) \quad (4)$$

Probabilitas pc_j dapat diestimasi dengan menghitung jumlah dokumen *training* pada setiap kategori cj . Sedangkan untuk menghitung distribusi $p(X|cj)$ akan sulit karena jumlah term menjadi sangat besar. Hal ini disebabkan jumlah term

tersebut sama dengan jumlah semua kombinasi posisi kata dikalikan dengan jumlah kategori yang akan diklasifikasikan. Dengan pendekatan Naïve Bayes yang mengasumsikan bahwa setiap kata dalam setiap kategori adalah tidak bergantung satu sama lain, maka perhitungan dapat lebih disederhanakan dan dapat dituliskan sebagai berikut :

$$p(X|_{c_j}) = II_{i+1}^{n} p(w_i \vee c_j) \quad (5)$$

Dengan menggunakan persamaan (2), maka persamaan (5) dapat dituliskan menjadi :

disajikan pada persamaan (1).

$$C_{MAP} = argmax\, p(c_j) II_{i=1}^{n} p(w_i \vee c_j) \quad (6)$$

Nilai p (c_J)dan p$(w_i \vee c_j)$ dihitung pada saat proses training dimana persamaannya adalah sebagai berikut :

$$P(c_j) = ¿\ \text{docs j} \vee \frac{¿}{¿\ \text{contoh} \vee ¿\ ¿}\ ¿ \quad (7)$$

$$P(w_i|c_j¿ = \frac{1+n_i}{|C|+n(kosakata)} \quad (8)$$

$p(w_i \vee c_j)$ = probabilitas kata w_i pada kategori c_j

|docs j| = jumlah dokumen pada kategori j

|contoh| = jumlah seluruh dokumen sampel yang digunakan dalam proses *training*

n_i = frekuensi kemunculan kata wi pada kategori cj

|C| = jumlah semua kata pada kategori cj

n(kosakata) = jumlah kata yang unik pada semua data training

III. METODE PENELITIAN

Pelaksanaan penelitian ini melibatkan beberapa sistematika penelitian dari pengumpulan data, *text preprocessing, feature selection,* klasifikasi data (*text mining*) hingga evaluasi hasil.

A. Pengumpulan Data

Data yang digunakan bersumber dari PT. PLN (Persero) WS2JB Area Palembang. Data tersebut merupakan data laporan pengaduan gangguan *response time* SLA.

B. Text Processing

Sebelum dilakukan *text processing*, data akan diolah dengan data mining terlebih dahulu, agar siap diolah pada *text mining*. Adapun proses yang dilakukan adalah *cleaning data* dan *transformation data*.

- *Cleaning Data*

  Data yang diperoleh masih belum sepenuhnya dapat diproses karena ada *field* kosong ataupun hilang. Maka harus dilakukan *cleaning data* untuk menghapus *field* kosong atau hilang tersebut. Sehingga diperoleh 12.881 *record* data dari sebelumnya 24.009 *record* data.

- *Transformation Data*

  Pada tahap ini format data diubah agar dapat digunakan untuk proses selanjutnya. Format data yang diubah adalah *record* durasi *recovery time*. *Record* ini diubah kedalam bentuk menit. Selain itu dibuat pula atribut klasifikasi yang berisi macam-macam klasifikasi dari keterangan laporan, klasifikasi ini terbagi menjadi 3, yaitu Dibawah 180 menit, Kurang dari 1 hari, dan Lebih dari 1 hari. *Transformation* ini dilakukan menggunakan *microsoft excel* dengan logika *IF ELSE*.

Text Preprocessing merupakan cara pemecahan teks ke dalam kata, yang mana kata tersebut akan menjadi kata kunci. Serangkaian langkah yang masuk ke dalam preprocessing diataranya *stemming*. Proses ini akan menghilangkan imbuhan pada kata sehingga tersisa kata dasar atau *root word*. Setelah didapatkan teks yang hanya tersusun dari kata dasar selanjutnya, dilakukan penghapusan kata-kata yang tidak diperlukan. Kata-kata yang tidak diperlukan ini biasa disebut

kata sambung atau *stopword*, oleh sebab itu proses ini disebut *stopword removal*. Hasil dari proses ini adalah teks yang terdiri dari kata kunci. Kemudian dilakukan tokenisasi. Proses ini akan mentransformasikan bentuk teks / string menjadi token-token, satu kata satu token. Selain itu pada proses ini juga terjadi penghapusan tanda baca sehingga hanya menghasilkan token yang berisi kata kunci. Berikut beberapa token yang dihasilkan:

"app" "appid" "kemarin" "meter"

"periksa" "terjadi"

C. Feature Selection

Pada tahapan ini akan dihitung *term frequency* (TF) dari setiap token. Nilai TF digunakan sebagai bobot untuk setiap kata kunci terhadap suatu dokumen (dalam hal ini keterangan laporan). Kemudian dihitung pula nilai DF (*document frequency*). DF dari sebuah fitur menunjukan jumlah dokumen dimana sebuah kata kunci tersebut muncul, seperti sebagai berikut:

Docs	app	appid	kemarin	meter	periksa	terjadi
1	0	0	0	0	0	1
2	1	1	1	1	2	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	1	0	0	1	2	0
8	0	0	0	1	0	0
9	0	0	0	0	0	0
10	0	0	0	1	0	0

Dalam hal ini Docs menunjukkan nomor dari *record* keterangan laporan.

D. Text Mining/ Pattern Discovery

Algoritma yang digunakan untuk mengolah *text mining* dalam penelitian ini adalah *Naive Bayes Classifier*. Dalam metode *Naive Bayes Classifier* dilakukan proses pengklasifikasian teks berdasarkan data *training* yang sebelumnya. Penelitian terkait *naive bayes classifiers* melalui beberapa tahap yakni *prior probability, conditional probability*, pemilihan kategori akan

ditentukan melalui nilai maksimum kategori terpilih [3].

IV. HASIL

Berdasarkan Gambar. I diketahui bahwa ada 2.487 keterangan laporan diklasifikasikan kedalam *recovery time* dibawah 180 menit, yang mana 446 teridentifikasi kedalam klasifikasi *recovery time* kurang dari 1 hari dan 14 keterangan laporan teridentifikasi kedalam *recovery time* lebih dari 1 hari.

Berdasarkan training selanjutnya, dilakukan penambahan estimator laplace=1. Sehingga diperoleh hasil pada gambar 2.

Berdasarkan gambar 1 dan 2, laporan yang teklasifikasi dibawah 180 menit ada 64,4% , 34,8% untuk laporan yang terklasifikasi kurang dari 1 hari dan 0,08% untuk laporan yang terklasifikasi lebih dari 1 hari.

V. KESIMPULAN

1. Jumlah kategori yang ada mempengaruhi kinerja klasifikasi teks menggunakan metode Naive bayes. Tingkat kemiripan diantara kategori satu dengan yang lain mempengaruhi tingkat akurasi klasifikasi teks. Jika tingkat kemiripan diantara dua kategori tinggi, maka akan sulit membedakan kedua kategori tersebut sehingga tingkat akurasi klasifikasi teks akan menurun.
2. Penggunaan *stopwords* dan *stemming* dapat meningkatkan tingkat akurasi klasifikasi teks.
3. Laporan yang teklasifikasi dibawah180 menit ada 64,4% , 34,8% untuk laporan yang terklasifikasi kurang dari 1 hari dan 0,08% untuk laporan yang terklasifikasi lebih dari 1 hari.

	Actual			
Predicted	Dibawah 180 menit	Kurang dari 1 hari	Lebih dari 1 hari	Row Total
Dibawah 180 menit	2020	1018	25	3063
	0.812	0.757	0.781	
Kurang dari 1 hari	424	312	7	743
	0.170	0.232	0.219	
Lebih dari 1 hari	43	15	0	58
	0.017	0.011	0.000	
Column Total	2487	1345	32	3864
	0.644	0.348	0.008	

Gambar 1: Hasil NBC

predicted	actual Dibawah 180 menit	Kurang dari 1 hari	Lebih dari 1 hari	Row Total
Dibawah 180 menit	2027	1014	24	3065
	0.661	0.331	0.008	0.793
	0.815	0.754	0.750	
Kurang dari 1 hari	446	326	8	780
	0.572	0.418	0.010	0.202
	0.179	0.242	0.250	
Lebih dari 1 hari	14	5	0	19
	0.737	0.263	0.000	0.005
	0.006	0.004	0.000	
Column Total	2487	1345	32	3864
	0.644	0.348	0.008	

Gambar 2: Hasil NBC dengan laplace 1

REFERENCES

[1] Aggarwal,C., and C. Zhai. 2012.Mining Text Data Chapter A Survey of Text Classification Algorithms. London : Kluwer Academics Publisher.

[2] Feldman, R., & Sanger, J. (2006). *The Text Mining Handbook*. https://doi.org/10.1017/CBO9780511546914

[3] Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*.

Elsevier (Vol. 12). https://doi.org/10.1007/978-3-642-19721-5

[4] Larose, D. T. (2005). *Discovering knowledge in data: an introduction to data mining*. *Statistics* (Vol. 1st). https://doi.org/10.1016/j.cll.2007.10.008