KLASIFIKASI PENGADUAN MASYARAKAT MENGGUNAKAN *NAIVE BAYES* BERBASIS SELEKSI ATRIBUT *INFORMATION GAIN*

Alter Lasarudin[1], Purwanto[2]
[12]Pasca Sarjana Teknik Informatika Universitas Dian Nuswantoro

ABSTRACT

The development of information every day is increasing. Public complaint is one form of information on the Internet is growing each day according to the number of people who make a complaint. In the management of complaints frequent errors in aduannya groupings so as to make the admin must work longer to perform grouping or classification of complaints. Such information becomes a media which is used for data mining research. One of the functions of data mining is classification. Naïve Bayes is one of the methods used for classification, one for the classification of documents or text. The classification is very useful for grouping data or documents by category. This will simplify the user data or documents in the search process. This research was conducted by applying the method Naïve Bayes for classification societies complaint data and algorithms Information Gain for the selection of attributes in order to improve the accuracy of the classification of public complaints. The test results by using 150 training data and testing the data 60 Naïve Bayes algorithm using attribute selection results without accuracy is 63.33%. whereas on testing Naïve Bayes algorithm using Information Gain attribute selection with the same data results are increasing even with k = 5. The best accuracy results found in this study was 86.67% using the selection attribute by 55.

Keywords: Naïve Bayes, Information Gain, Classification, Public Complaints

1. PENDAHULUAN

Peningkatan jumlah informasi saat ini di internet sangatlah pesat. Hal ini dibuktikan dengan 550 triliun lebih dokumen dan 7.3 juta setiap hari halaman web baru [1]. Informasi tersebut bukan hanya dalam bentuk teks tapi juga dalam bentuk dokumen dengan berbagai format seperti pdf, Ms. Word, Ms. Excel, Ms. Powerpoint dan txt. Peningkatan informasi ini menjadi sebuah isu yang menarik untuk dijadikan riset tentang teks mining. Teks mining dapat menjadi solusi dalam menyelesaikan masalah seperti memproses, mengorganisasi, dan menganalisis teks yang tidak terstruktur dengan jumlah yang besar. Tujuan utama teks mining adalah peringkasan merupakan proses mengelompokkan dokumen sesuai kategorinya. Manfaat klasifikasi dokumen sangatlah besar karena jumlah dokumen setiap hari semakin bertambah sehingga peran dari klasifikasi dokumen sangatlah penting.

Pengaduan masyarakat merupakan salah satu bentuk informasi yang ada di internet yang semakin hari semakin bertambah sesuai dengan jumlah orang yang melakukan pengaduan. Pada PNPM Mandiri Perkotaan yang saat ini lebih dikenal dengan nama Kota Tanpa Kumuh (KOTAKU), pengaduan masyarakatnya lebih dikenal dengan Pengelolaan Pengaduan Masyarakat (PPM). Pada PPM tersebut terdapat beberapa kategori antara lain kritik, saran, pertanyaan, pelanggaran mekanisme dan prosedur, penyimpangan dana, adanya intervensi negatif, perubahan kebijakan, kode etik/kinerja pelaku, dan Force Majeur. Rekapan data aduan pada bulan Agustus 2016 berjumlah 1812 data aduan yang terbagi dalam 7 dari 9 kategori yang ada. Hal ini dikarenakan jenis aduan yang dilakukan oleh masyarakat lebih dominan pada kategori pertanyaan, saran dan kritik. Untuk lebih jelasnya data aduan yang masuk pada database PPM Kota Tanpa Kumuh dapat dilihat pada tabel berikut.

Tabel 1. Data Aduan Masyarakat [2]

Kategori	Jumlah Aduan	Persentase
Kritik	69	3.81%
Saran	141	7.78%
Pertanyaan	1574	86.87%
Pelanggaran Mekanisme dan Prosedur	1	0.06%
Penyimpangan Dana	0	0.00%
Adanya Intervensi Negatif	0	0.00%
Perubahan kebijakan	3	0.17%
Kode Etik / Kinerja Pelaku	22	1.21%
Force Majeur	2	0.11%
Total Aduan	**1812**	

Dalam pengelolaan pengaduan sering terjadi kesalahan dalam pengelompokan aduannya sehingga membuat admin harus bekerja lagi untuk melakukan pengelompokan atau pengklasifikasian aduan. Berdasarkan hasil pengamatan peneliti bahwa persentasi kesalahan dalam data pengaduan masyarakat ± 30%.

Menurut Chen, Huang, Tian & Qu, 2009 dalam Lila Dini Utami [3], bahwa untuk klasifikasi teks, algoritma *Naïve Bayes* banyak digunakan dikarenakan algoritma ini sangat sederhana dan efisien serta dengan menggunakan seleksi atribut hasilnya sangat sensitif. Menurut Xhemali, et al. 2009 dalam Ivan Jaya [4], bahwa algoritma *Naïve Bayes* adalah algoritma klasifikasi yang setiap atributnya bersifat indenpenden. Atribut-atribut pada data kemungkinan memiliki nilai tidak relevan untuk melakukan tugas mining dan jika atribut tersebut disertakan dalam proses mining dapat mengacaukan tugas algoritma data mining. Oleh karena itu perlu dilakukan proses seleksi atribut yang memiliki nilai tidak relevan atau bahkan berlebihan [4].

Untuk melakukan seleksi atribut dapat menggunakan metode penyaringan, atibut diurutkan sesuai dengan peringkat berdasarkan evaluasi dengan kriteria tertentu. Informasi Gain Salah merupakan salah satu satu algoritma seleksi atribut dengan metode penyaringan. Semakin besar nilai informasi gain maka semakin penting atribut tersebut [4].

Yulia Sulistyaningsih dkk [5], dalam penelitiannya dengan menggunakan metode C4.5 dan 72 data aduan yang dijadikan data tes menunjukkan bahwa akurasi, precision, recall dan f-measure untuk data tes dengan persebaran data kelas yang tidak merata berturut-turut 68,06%, 72,4%, 69,8% dan 68%. Sedangkan nilai akurasi, precision, recall dan f-measure untuk data latih dengan persebaran data kelas yang merata berturut-turut 76,39%, 76,9%, 76,4% dan 75,4%.

Cahyono Darujati dan Agustinus Bimo Gumelar [6] melakukan penerapan metode Naive Bayes Classifier untuk mengklasifikasi berita yang bersumber dari situs web dan hasil nilai akurasinya lebih dari 87 % untuk data latih yang besar yaitu 100 artikel.

Betna Nurina Sari [7], menyimpulkan bahwa hasil eksperimen dengan menggunakan teknik pemilihan fitur *Information Gain*, performa algoritma klasifikasi machine learning (J48, Random Forest, MLP, SVM (SMO), dan *Naïve Bayes*) untuk memprediksi performa akademik siswa pada mata pelajaran Matematika dapat ditingkatkan.

Akhmad Pandhu Wijaya dan Heru Agus Santoso [8], penelitian ini bertujuan untuk mengklasifikasi dokumen bahasa Indonesia untuk mengidentifikasi konten e-Government pada situs website Jawa Tengah dengan menggunakan algoritma *Naïve Bayes* Clasification (NBC) dan pembobotan fitur dengan metode TF-IDF untuk menghasilkan nilai dan akurasi yang baik. Hasil dari penelitian klasifikasi dokumen menggunakan algoritma NBC dengan data tranning sebanyak 260 dokumen politik dan 222 dokumen ekonomi menggunakan 40 data latih dengan akurasi sebesar 85%.

Lila Dini Utami [3], mengatakan bahwa penelitiannya menggunakan Algoritma *Information Gain* untuk seleksi fiturnya dan adaboost untuk pengurangan bias dalam meningkatkan akurasi Algoritma *Naïve Bayes*. Hasil penelitian adalah untuk mengklasifikasi teks ke bentuk positif atau negatif review restoran dengan hasil akurasinya mengalami peningkatan dari 73.00% menjadi 81.50%.

Amir Hamzah [9], dalam penelitiannya bahwa metode *Naïve Bayes* memiliki kelemahan mengansumsi independensi fitur kata. Tujuan penelitian ini adalah untuk mengkaji akurasi algoritma *Naïve Bayes* untuk pengkategorian teks berita dan teks akademik. Hasil penelitian akurasi pada dokumen berita dengan 1000 dokumen mencapai 91% sedangkan untuk dokumen akademik akurasinya adalah 82%.

Andre L.F. Alves, dkk [10], dengan menggunakan dataset sekitar 300.000 tweet berbahasa Portugis mengenai tema FIFA CUP yang berlangsung di Brasil pada tahun 2013. Data tersebut diklasifikasikan dengan menggunakan SVM dan *Naïve Bayes*. Penelitian ini membandingkan hasil akurasi algoritma SVM dengan *Naïve Bayes* untuk mengklasifikasi sentiment kata pada tweet. Hasil yang diperoleh oleh SVM, sentimen diklasifikasikan mengindikasikan F-Measure = 0,873 dan akurasi = 80,0% untuk deteksi sentimen polaritas. Sedangkan *Naïve Bayes*, sentimen klasifikasi disajikan F-Measure = 0,791 dan akurasi = 72,7%.

Tujuan dari penelitian ini adalah meningkatkan akurasi klasifikasi pengaduan masyarakat menggunakan algoritma *Naïve Bayes* dan algoritma *Information Gain* untuk seleksi atribut dalam mengklasifikasi pengaduan masyarakat pada tiga kategori pengaduan yaitu kategori saran, kritik dan pertanyaan.

Manfaat dari penelitian ini diharapkan dapat berguna sebagai sumbangan pemikiran khususnya di bidang teks mining. Sebagai referensi bagi peneliti selanjutnya yang ingin meneliti tentang teks mining, khususnya tetang pengklasifikasian dokumen atau teks dengan metode *Naïve Bayes* dan algoritma *Information Gain*.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terkait

Penelitian yang dilakukan oleh Yulia Sulistyaningsih dkk [5] menggunakan algoritma klasifikasi pohon keputusan (C4.5) untuk mengklasifikasi pengaduan masyarakat. Penelitian ini menggunakan 6 kategori pengaduan dengan masing-masing kategori memiliki 40 data pengaduan masyarakat. Proses perbandingan pohon dilakukan dengan enam percobaan menggunakan partisi data latih banding data uji 70:30.

Tabel 2. Perbandingan Kinerja Pohon Keputusan [5]

Percobaan	Akurasi (%)		
	C4.5	RandomForest	RandomTree
1	59,51	56,59	42,44
2	57,56	59,51	41,44
3	55,76	55,15	48,48
4	58,79	55,15	41
5	62,50	61,11	38,9
6	65,28	55,56	51,39

Penelitian yang dilakukan oleh Cahyono Darujati dan Agustinus Bimo Gumelar [6] melakukan penerapan metode *Naive Bayes Classifier* (NBC) untuk mengklasifikasi berita yang bersumber dari situs web. Himpunan data yang digunakan pada penilitian ini adalah kumpulan artikel yang diambil dari majalah CHIP yang dibagi dalam kelas-kelasnya yaitu Komputer Teknologi, Kesehatan, dan Berita dengan jumlah total data ± 3000 artikel. Hasil penelitian menyatakan bahwa akurasi pada klasifikasi berita

dengan metode *Naïve Bayes* sangat baik. Hal ini dibuktikan dengan data uji yang diambil dari website menghasilkan akurasi 87% dengan data latih 100 artikel.

Betha Nurina Sari [7], meneliti tentang implementasi seleksi fitur *Information Gain* pada algoritma klasifikasi untuk mengidentifikasi pengaruh performa akurasi pengklasifikasian akademik siswa. Data yang digunakan adalah data 395 siswa untuk evaluasi nilai akhir pada mata pelajaran Matematika. Pada hasil eksperimen menggunakan semua atribut dataset dengan menggunakan lima algoritma klasifikasi *machine learning* untuk prediksi performa akademik siswa seperti tampak pada tabel di bawah ini.

Tabel 3. Akurasi Sebelum Seleksi Fitur

Algoritma Klasifikasi	**Akurasi**	
	Skenario A	**Skenario B**
Decision Tree (J48)	88.58	100
Random Forest	90.43	93.04
Neural Network (MLP)	88.15	73.65
SVM (SMO)	89.14	66.2
Naïve Bayes	85.67	80.46

2.2. Landasan Teori

2.2.1 *Data Mining*

Data mining merupakan proses komputasi untuk menganalisis data dalam jumlah besar dengan mengekstrak pola dan informasi yang berguna. Istilah data mining memiliki beberapa padanan, seperti *knowlodge discovery* ataupun *pattern recognition*. Jadi tujuan utama data mining adalah untuk menemukan, menggali, atau menambang pengetahuan dari data atau informasi yang kita miliki.

Terdapat enam fungsi dalam data mining, yatu fungsi deskripsi, fungsi estimasi, fungsi prediksi, fungsi klasifikasi, fungsi pengelompokan dan fungsi asosiasi. Keenam fungsi data mining tersebut dapat kelompokkan menjadi 2 fungsi lagi, yaitu fungsi minor atau fungsi tambahan dan fungsi mayar atau fungsi utama [16]. Fungsi minor meliputi tiga fungsi pertama, yaitu deskripsi, estimasi, dan prediksi. Sedangkan fungsi mayor yaitu klasifikasi, pengelempokkan dan asosiasi.

2.2.2 *Text Mining*

Secara umum pekerjaan teks mining mirip dengan pekerjaan data mining, yaitu penggalian prediktif dan penggalian deskriptif. Teks Mining merupakan proses ekstraksi pola informasi dan pengetahuan yang berguna dari sejumlah sumber data teks, seperti dokumen Word, PDF, kutipan teks, dll [10]. Proses yang sering dilakukan oleh teks mining adalah perangkuman otomatis, kategorisasi dokumen, penggugusan teks, dan lain-lain. Hal utama yang membedakan antara teks mining dan data mining adalah pada data masukkannya. Daya masukkan pada teks mining adalah data yang tidak terstruktur sedangkan data masukkan pada data mining adalah data yang sudah terstruktur.

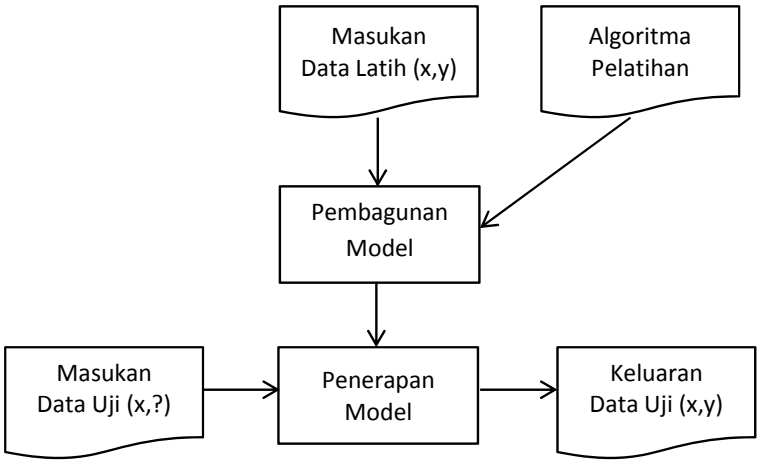
Secara umum terdapat dua tahapan dalam melakukan teks mining. Tahapan tersebut adalah tahapan *text preprocessing* dan tahapan *feature selection*. Pada tahapan *text preprocessing* terdapat tahapan-tahapan lagi yaitu (1) tahap *case folding* yaitu merubah karakter huruf capital menjadi huruf kecil, (2) tahap *tokenize* yaitu proses peruraian kalimat menjadi kata-kata dan menghilangkan delimiter seperti tanda titik (.), koma (,) spasi dan karakter angka yang ada pada kata tersebut. Sedangkan pada tahap feature selection terdapat proses (1) *stopword removal* yaitu proses menghilangkan kata-kata dasar seperti kata “di”, “yang”, “dan”, (2) *stemming* yaitu proses menghilangkan imbuhan-imbuhan dan hasilnya adalah kata dasar.

Penggunaan dan penelitian tentang teks mining sudah banyak dilakukan seiring dengan perkembangan jumlah data teks atau dokumen yang mudah didapat dari berbagai media, seperti sosial media, website dan

lain sebagainya. Teks mining memiliki tujuh area praktek, yaitu (1) pencarian dan perolehan informasi; (2) pengelompokan dokumen; (3) klasifikasi dokumen; (4) web mining; (5) ekstraksi informasi; (6) *Natural Language Processing* (NLP); dan (7) ekstraksi konsep.

2.2.3 Klasifikasi

Klasifikasi dapat diartikan sebagai suatu pekerjaan yang melakukan pelatihan atau pembelajaran terhadap fungsi target *f* yang memetakan setiap vector (set fitur) *x* ke dalam satu dari sejumlah label kelas *y* yang tersedia. Pekerjaan pelatihan tersebut akan menghasilkan suatu model yang kemudian disimpan sebagai memori [11].



Gambar 1. Proses Pekerjaan Klasifikasi [11]

2.2.4 *Naïve Bayes* Classifier (NBC)

Bayes merupakan teknik prediksi berbasis probabilitas sederhana yang berdasar pada penerapan teorema Bayes dengan asumsi indenpedensi (ketidakketergantungan) yang kuat (naïf). Bayes dikenal juga dengan istilah *Naïve Bayes*. Dalam *Naïve Bayes* maksud independensi yang kuat pada fitur adalah sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidak adanua fitur lain dalam data yang sama.

Formulasi *Naïve Bayes* untuk klasifikasi teks adalah sebagai berikut :

$$p(c_i)\frac{f_d(c_i)}{|D|} \quad \text{.....(1)}$$

$f_d(c_i)$ adalah jumlah dokumen yang memiliki kategori c_i

|D| adalah jumlah seluruh *training* dokumen

untuk mencari nilai probabilitas dari setiap kategori atau *class*.

$$p(w_k|c_i) = \frac{n_k+1}{n+|vocabulary|} \quad \text{.....(2)}$$

n_k adalah nilai kemunculan kata pada kategori c_i

n adalah jumlah keseluruhan kata pada kategori c_i

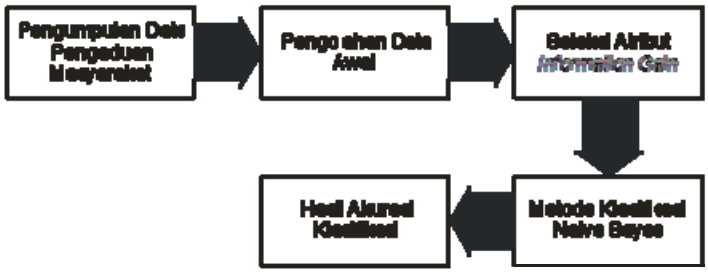
|vocabulary| adalah jumlah keseluruhan kata

2.2.5 *Information Gain*

Information Gain merupakan salah satu algoritma seleksi atribut yang digunakan untuk meningkatkan akurasi dalam pengklasifikasian. Sebelum menghitung nilai *Information Gain* tentukan dulu *entropy*, karena setiap objek yang diklasifikasi harus diuji dulu nilai entropinya.

3. METODE PENELITIAN

3.1. Tahapan Klasifikasi Pengaduan Masyarakat



Gambar 2. Diagram Alur Klasifikasi

3.2. Pengumpulan Data

Dataset yang digunakan pada penelitian ini adalah data pengaduan masyarakat yang terdiri dari tiga kategori yaitu saran, kritik dan pertanyaan. Dataset diambil dari website www.p2kp.org pada tanggal 20 September 2016. Jumlah dataset yang digunakan sebanyak 210 data aduan yang di bagi ke dalam tiga kategori dengan masing-masing 70 data aduan. Dataset diklasifikasi berdasarkan isi aduan setiap kategori.

3.3. Pengolahan Data Awal

3.3.1 Tokenisasi

Tokenisasi merupakan proses pemotongan teks menjadi beberapa bagian yang disebut token. Pada tahap ini bertujuan untuk memisahkan kata per kata dari kalimat setiap aduan. Sebelum melakukan tahapan tokenisasi, dilakukan tahap merubah huruf kapital menjadi huruf kecil atau yang dikenal dengan istilah *case folding*. *Case folding* adalah merubah capital menjadi huruf kecil. Setelah melakukan tahap case folding dilanjutkan tahap normalisasi kata, yaitu proses perbaikan kata-kata yang salah eja atau disingkat dalam bentuk tertentu. Perbaikan kata dilakukan secara manual untuk menghindari jumlah perhitungan dimensi kata yang melebar. Perhitungan dimensi kata akan melebar jika kata yang salah eja atau disingkat tidak diubah, karena kata tersebut sebenarnya mempunyai maksud dan arti yang sama tetapi akan dianggap sebagai entitas yang berbeda, Setelah melakukan normalisasi kata, selanjutnya dilakukan proses tokenisasi. Tahapan ini juga bertujuan untuk menghilangkan tanda baca yang tidak diperlukan seperti “,”, “.”, “/”, “?”, angka dan lain sebagainya.

3.3.2 Stopword

Setelah melakukan tekonisasi, tahap selanjutnya adalah membuang kata-kata yang dianggap tidak penting atau yang dikenal dengan *stopword.* Caranya adalah dengan mencocokkan data *stopword* atau kata dasar yang ada pada kata bahasa Indonesia. Seperti kata “yang”, “akan”, dan lain sebagainya.

3.3.3 Stemming

Stemming adalah teknik untuk menemukan kata dasar dari sebuah kata yang mengandung kata berimbuhan. Misalnya kata “penggunaan”, “menggunakan”, “digunakan” dan “berguna” memiliki bentuk dasar yang sama yaitu “guna”. Pada penelitian ini proses stemming yang digunakan adalah stemming Sastrawi. Proses stemming oleh Sastrawi sangat bergantung pada kamus kata dasar. Sastrawi menggunakan kamus kata dasar dari kateglo.com dengan sedikit perubahan [11].

3.3.4 Term Weighting

Pada tahap ini peneliti menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). TF-IDF adalah metode pembobotan yang mengaitkan antara *term frequency* (TF) dan *inverse document frequency* (IDF) [8]. Skema persamaan TF-IDF ditunjukkan oleh persamaan berikut:

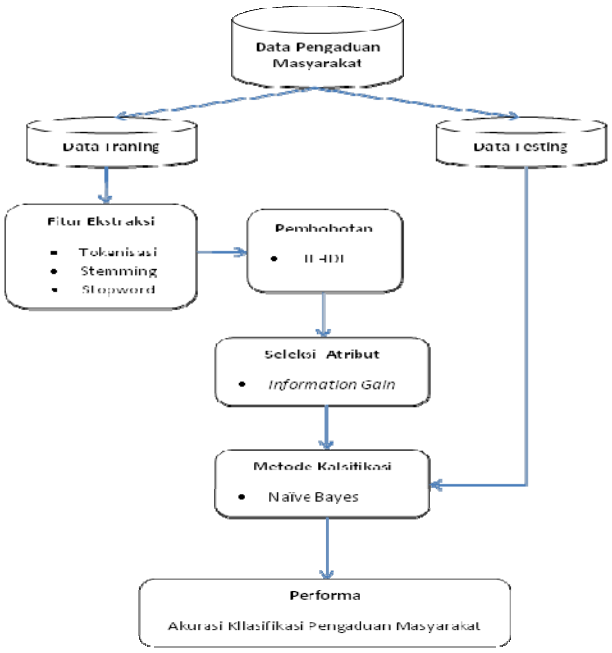
$$tfidf(w) = tf \times log\frac{n}{df(w)} \quad \text{..... (3)}$$

Keterangan:
tf (w) = *Term frequency* (jumlah kemunculan suatu kata dalam suatu dokumen)
df (w) = *Document frequency* (jumlah dokumen yang mengandung suatu kata)
N = Jumlah dokumen

Term Frequency (TF) adalah jumlah kemunculan suatu kata pada satu dokumen atau aduan. *Document Frequency* (DF) adalah jumlah dokumen atau aduan yang mengandung suatu kata. *Inverse Document Frequency* (IDF) adalah hasil bagi jumlah dokumen dengan *Document Frequency*. Untuk mendapatkan nilai *document frequency* yaitu dengan menjumlahkan kemunculan suatu kata pada semua dokumen.

3.4. Metode Ekperimen

Penelitian ini menggunakan metode *Naïve Bayes* untuk mengklasifikasi teks dan algoritma *Information Gain* untuk seleksi atributnya. Atribut-atribut yang digunakan adalah kata-kata yang telah di *stemming* dengan *stemming* Bahasa Indonesia Sastrawi. Setelah dilakukan seleksi atribut dengan *Information Gain* selanjutnya dilakukan proses pengklasifikasian dengan menggunakan algoritma *Naïve Bayes*. Teknik seleksi atribut diterapkan dengan cara mengitung nilai fitur yang memiliki nilai tertinggi.



Gambar 3. Model Eksperimen

3.5. Evaluasi dan Hasil Validasi

Penelitian klasifikasi untuk evaluasi hasil pengujian menggunakan cross validation untuk mendapatkan hasil *accuracy*, *precision* dan *recall* terbaik. Pengujian *cross validation* dilakukan sampai 10-fold.

Accuracy adalah tingkat dari dokumen yang benar diidentifikasi, sedangkan *precesion* adalah perbandingan jumlah data yang sesuai dengan data yang diminta, sedangkan *recall* adalah tingkat keberhasilan sistem dalam menemukan kembali informasi.

4. HASIL PENELITIAN

4.1. Hasil Pengujian Algoritma *Naïve Bayes* dan *Information Gain*

Dengan menggunakan formulasi *Naïve Bayes* peneliti melakukan klasifikasi dengan menggunakan 6 data *training* dan 1 data *testing*.

Tabel 4. Data Training dan Testing

Aduan	Kategori
giat monitoring perintah pusat tidak seluruh maksimal giat desa	Kritik
rencana terus tidak realisasi	Kritik
upk perlu tahan punya mampu kelola dana gulir administrasi	Saran
temu bentuk pokja pilih bkm tunda dulu hadir koordinator bkm	Saran
kapan timm pokjanis propinsi bentuk apa kendala tim sebut bentuk	Pertanyaan
apakah program kotaku aspek laksana lapang atas	Pertanyaan
tidak punya dana gulir	?

Berdasarkan data yang ada pada Tabel 4 tampak setiap kategori memiliki 2 dokumen aduan. Nilai probabilitas pada masing-masing kategori adalah sebagai berikut :

P(Kritik) = $\frac{2}{6} = 0{,}333$

P(Saran) = $\frac{2}{6} = 0{,}333$

P(Pertanyaan) = $\frac{2}{6} = 0{,}333$

Terdapat40 kata unik yang ditemukan pada 6 dokumen aduan data *training* masing-masing kategori atau |vocabulary| = 40.

Dengan menggunakan hasil kata unik pada masing-masing kategori dilakukan perhitungan probabilitas setiap kata yang ada pada masing-masing kategori. Untuk menghitung nilai probabilitas setiap kata pada masing-masing kategori. Sebagai contoh peneliti menghitung nilai probabilitas kata “bentuk” pada kategori saran dan pertanyaan. pada kategori saran nilai nk kata bentuk = 1 dan pada kategori pertanyaan = 2.

P(bentuk|saran) = $\frac{1+1}{16+40} = \frac{2}{56} = 0.0357$

P(bentuk|pertanyaan) = $\frac{2+1}{18+40} = \frac{3}{58} = 0.0545$

Dengan melihat hasil perhitungan diatas maka dapat simpulkan bahwa kata bentuk memiliki probabilitas tertinggi pada kategori pertanyaan.

Setelah mendapatkan nilai probabilitas setiap kata pada masing-masing kategori kemudian dicari kategori dari data *testing* berdasarkan nilai probabilitas tertinggi dari masing-masing kategori dengan cara menjumlahkan nilai probabilitas setiap kata yang ada pada data *training* disesuaikan dengan kata yang ada pada data *testing*.

Hasi Probabilitas pada kategori kritik adalah :

P(kritik) = P(kritik) + P(tidak) + P(punya) + P(dana) + P(gulir)
= 0.333 + 0.0566 + 0.0188 + 0.0188 + 0.0188
= **0.446**

Hasi Probabilitas pada kategori saran adalah :

P(saran) = P(saran) + P(tidak) + P(punya) + P(dana) + P(gulir)
= 0.333 + 0.0178 + 0.357 + 0.357 + 0.357

= **0.458**

Hasi Probabilitas pada kategori pertanyaan adalah :

P(pertanyaan) = P(pertanyaan) + P(tidak) + P(punya) + P(dana) + P(gulir)

= 0.333 + 0.0181 + 0.0181 + 0.0181 + 0.0181

= **0.0727**

Berdasarkan hasil probabilitas pada masing-masing kategori, probabilitas tertinggi ada pada kategori saran yaitu 0.458. dengan demikian data *testing* pada tabel 4.5 termasuk pada kategori saran.

4.2. Hasil Pengujian Algoritma *Naïve Bayes* dengan Seleksi Atribut *Information Gain*

Seleksi atibut dengan algoritma *Information Gain* dihitung untuk mencari nilai entropy dan nilai gain.

Tabel 5. Pengujian *Information Gain*

Dok ke	Atribut						Class
	bentuk	apakah	dana	giat	gulir	tidak	
1	0	0	0	2	0	1	Kritik
2	0	0	0	0	0	1	Kritik
3	0	0	1	0	1	0	Kritik
4	0	0	0	0	0	0	Kritik
5	0	0	1	0	1	0	Kritik
6	0	0	0	0	0	1	Saran
7	0	0	0	0	0	0	Saran
8	1	0	0	0	0	0	Saran
9	0	0	1	0	1	0	Saran
10	1	0	0	0	0	0	Saran
11	2	0	0	0	0	0	Pertanyaan
12	0	1	0	0	0	0	Pertanyaan
13	0	1	0	0	0	0	Pertanyaan
14	0	1	0	1	0	0	Pertanyaan

Dengan menggunakan data pada diatas dilakukan pengujian seleksi atribut dengan *Information Gain*. Sebelum mencari nilai entropy dan *Information Gain*, terlebih dahulu mencari nilai probabilitas dari masing-masing kategori. Nilai probabilitas tersebut nantinya akan digunakan untuk mencari nilai entropy.

Probabilitas Kategori/class:

Kritik : $= \frac{5}{14} = 0.357$

Saran : $= \frac{5}{14} = 0.357$

Pertanyaan : $= \frac{4}{14} = 0.285$

Entropy : = -0.357 log(0.357,2) – 0.357 log(0.357,2) – 0,285 log(0.285,2) = 1.577

Setelah mendapatkan nilai entropy pada kategori/class dilanjutkan dengan mencari nilai probabilitas setiap atribut terhadap kategori.

Probabilitas atribut bentuk :

P(bentuk=0) : kritik = 5/11 = 0.454; saran = 3/11 = 0.272; pertanyaan = 3/11 = 0.272

Entropy = 1.539

P(bentuk=1) : kritik = 0/2 = 0; saran = 2/2 = 1; pertanyaan = 0/2 = 0

Entropy = 0
P(bentuk=2) : kritik = 0/1 = 0; saran = 0/1 = 0; pertanyaan = 1/1 = 1
Entropy = 0
Information Gain = 1.577 – (((11/14)* 1.539)+((2/14)*0)+((1/14)*0)) = **0.368**

Probabilitas atribut apakah :
P(apakah=0) : kritik = 5/11 = 0.454; saran = 5/11 = 0.454; pertanyaan = 1/11 = 0.091
Entropy = 1.349
P(apakah =1) : kritik = 0/3 = 0; saran = 0/3 = 0; pertanyaan = 3/3 = 1
Entropy = 0
Information Gain = 1.577 – (((11/14)* 1.349)+((3/14)*0 = **0.517**

Probabilitas atribut dana :
P(dana=0) : kritik = 3/11 = 0.272; saran = 4/11 = 0.363; pertanyaan = 4/11 = 0.363
Entropy = 1.572
P (dana =1) : kritik = 2/3 = 0.666; saran = 1/3 = 0.333; pertanyaan = 0/3 = 0
Entropy = 0
Information Gain = 1.577 – (((11/14)*1.572)+((3/14)*0 = **0.342**

Probabilitas atribut giat :
P(giat=0) : kritik = 4/12 = 0.333; saran = 5/12 = 0.416; pertanyaan = 3/12 = 0.25
Entropy = 1.555
P (giat =1) : kritik 0/1 = 0; saran = 0/1 = 0; pertanyaan = 1/1= 1
Entropy = 0
P (giat =2) : kritik 1/1 = 1; saran = 0/1 = 0; pertanyaan = 0/1= 0
Entropy = 0
Information Gain = 1.577-(((11/14)*1.555)+((1/14)*0)+((1/14)*0 = **0.355**

Probabilitas atribut gulir :
P(gulir=0) : kritik = 2/11 = 0.181; saran = 4/11 = 0.363; pertanyaan = 4/11 = 0.363
Entropy = 1.508
P (gulir=1) : kritik 2/3 = 0.666; saran = 1/3 = 0.333; pertanyaan = 0/3= 0
Entropy = 0
Information Gain = 1.577 – (((11/14)* 1.508)+((3/14)*0 = **0.392**

Probabilitas atribut tidak :
P(tidak=0) : kritik = 3/11 = 0.272; saran = 4/11 = 0.363; pertanyaan = 4/11 = 0.363
Entropy = 1.572
P (tidak =1) : kritik 2/3 = 0.666; saran = 1/3 = 0.333; pertanyaan = 0/3= 0
Entropy = 0
Information Gain = 1.577 – (((11/14)* 1.572)+((3/14)*0 = **0.342**

Tabel 6. Hasil *Information Gain*

Atribut	IG
bentuk	0.368
apakah	0.517
dana	0.342
giat	0.355

gulir	0.392
tidak	0.342

Nilai *Information Gain* tertinggi ada pada kata atau atribut apakah kemudian kata gulir, bentuk, giat, dana, dan tidak. Berdasarkan nilai tersebut apabila kita akan melakukan seleksi atribut maka atribut-yang memiliki nilai-nilai tertinggi yang akan diambil.

4.3. Evaluasi dan Hasil Validasi

Penggujian dengan menggunakan tool rapidminer, penggujian algoritma *Naïve Bayes* untuk klasifikasi aduan tanpa menggunakan seleksi atribut *Information Gain*. Hasil akurasi yang didapatkan sebesar 63.33%.

Tabel 7. Hasil Akurasi Algoritma *Naïve Bayes* tanpa *Information Gain*

Accuracy: 63.33%				
	true Kritik	true Saran	true Pertanyaan	**class *precision***
pred. Kritik	26	8	7	**63.41%**
pred. Saran	10	38	12	**63.33%**
pred. Pertanyaan	14	4	31	**63.27%**
class recall	**52.00%**	**76.00%**	**62.00%**	

Berdasarkan tabel di atas dapat dilihat bahwa masih banyak kesalahan dalam pengklasifikasian aduan. Terutama pada kategori kritik, dari 50 data latih yang termasuk kategori kritik hanya 26 data, 10 data diprediksi saran dan 14 data diprediksi pertanyaan. Pada kategori saran 38 data yang di klasifikasi benar dari 50 data, 8 data di prediksi kritik dan 4 data diprediksi pertanyaan. sedangkan pada kategori pertanyaan yang diklasifikasikan benar ada 31 data, 7 data diprediksi kritik dan 12 data diprediksi saran.

Pada pengujian kedua peneliti melakukan pengujian algoritma *Naïve Bayes* dengan menggunakan seleksi atribut *Information Gain*. Penggujian dengan *Information Gain* dilakukan dengan menentukan jumlah atribut yang akan digunakan. Jumlah atribut ditentukan mulai dengan (*k*) = 5 sampai 100 dengan kelipatan 5 (k = 5, 10, 15, …, 100). Penentuan jumlah atribut ini dilakukan untuk mencari nilai akurasi terbaiknya berada pada jumlah seleksi atribut berapa.

Tabel 8. Hasil Akurasi Algoritma *Naïve Bayes* + IG untuk k=5

Accuracy:	**68.67%**			
	true Kritik	true Saran	true Pertanyaan	**class *precision***
pred. Kritik	47	41	2	**52.22%**
pred. Saran	3	9	1	**69.23%**
pred. Pertanyaan	0	0	47	**100.00%**
class recall	**94.00%**	**18.00%**	**94.00%**	

Pada percobaan pertama dengan menggunakan k = 5, tingkat keberhasilan sistem dalam mengklasifikasikan data sudah 94% pada kategori kritik dan pertanyaan, tetapi sangat kecil pada kategori saran yang hanya 18%. Pada kategori saran lebih banyak di prediksi kritik. Akurasi yang didapatkan adalah 68.67. Jika dibandingkan dengan pengujian pertama tanpa menggunakan seleksi atribut

Information Gain hasil akurasinya meningkat 5.34%.

Tabel 9. Hasil Akurasi Algoritma *Naïve Bayes* + IG untuk k=55

accuracy:	**86.67%**			
	true Kritik	true Saran	true Pertanyaan	class *precision*
pred. Kritik	44	5	3	84.62%
pred. Saran	2	42	3	89.36%
pred. Pertanyaan	4	3	44	86.27%
class recall	88.00%	84.00%	88.00%	

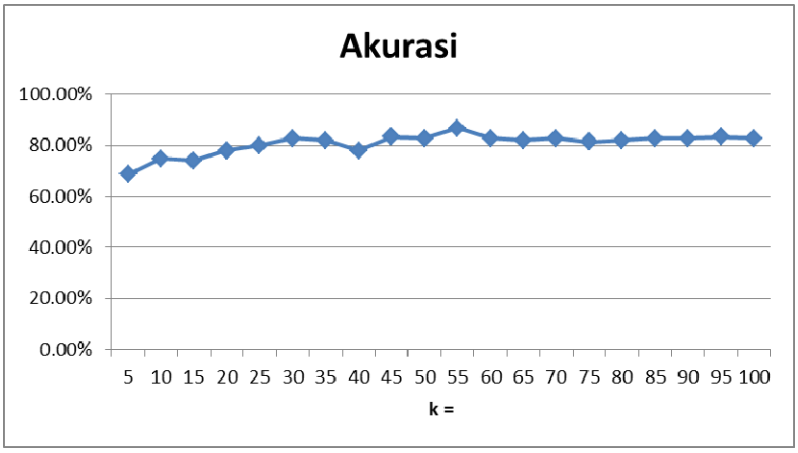
Pada percobaan selanjutnya dengan menggunakan k=55 hasil akurasi yang didapatkan meningkat sangat signifikan, yaitu sebesar 86.67%. Pada percobaan ini hasil klasifikasi setiap kategori sudah sangat baik walaupun masih ada beberapa kesalahan dalam pengklasifikasian. Recall untuk kritik yaitu 76%, saran, 84% dan pertanyaan 88%.

Berdasarkan hasil percobaan dengan menggunakan k = 5 sampai dengan 100 dengan kelipatan 5, hasil akurasi terbaik ditemukan pada pengujian k = 55 atribut terpilih dengan nilai akurasi sebesar 86,67%.

Tabel 10. Hasil Seleksi Atribut Terbaik *Naïve Bayes* + IG

Jumlah Atribut (k)	**Akurasi**	**Atribut yang Terpilih**
55	**86.67%**	administrasi, aktif, apakah, bagaimana, baik, baseline, bentuk, beri, buang, data, drainase, dulu, ekonomi, giat, gulir, hendak, informasi, jadi, jalan, kalau, kapan, kerja, kotaku, kumuh, kurang, laksana, lanjut, latih, lebih, libat, lokasi, lurah, masuk, masyarakat, minta, mohon, perlu, pilih, prioritas, program, progres, rtlh, sampah, segera, suatu, sulit, sumbat, tahun, tangan, tempat, tidak, titik, tunggak, ubah

Sedangkan untuk *precision* dan *recall* kategori kritik, saran dan pertanyaan pada percobaan dengan menggunakan k = 55 masing masing adalah untuk precision 84.62%, 89.36% dan 86.27% sedangkan untuk *recall* yaitu 88%, 84% dan 88%. Hal ini menyatakan bahwa dataset yang diprediksi pada masing-masing kategori hampir semuanya benar.



Gambar 4. Grafik Akurasi *Naïve Bayes* + IG

5. KESIMPULAN DAN SARAN

5.1. Kesimpulan

Hasil penelitian ini menggunakan data *training* sebanyak 150 dan data *testing* sebanyak 20 data pada masing-masing kategori kritik, saran dan pertanyaan. Hasil pengklasifikasian menggunakan algoritma *Naïve Bayes* tanpa menggunakan metode seleksi atribut *Information Gain* yang diuji menggunakan tool rapidminer mendapatkan akurasi sebesar 63.33%. Hasil *Precision* pada kategori kritik, saran dan pertanyaan masing masing adalah, 63.41%, 63.33% dan 63,27% sedangkan untuk *Recall* adalah 52%, 76% dan 62%. Untuk meningkatkan hasil akurasi klasifikasi data aduan dengan algoritma *Naïve Bayes* digunakan metode seleksi atribut *Information Gain*. Seleksi atribut dilakukan dengan menentukan jumlah atribut yang akan di pilih (k). Jumlah k yang digunakan adalah mulai dari k = 5 sampai dengan 100 dengan kelipatan 5. Hasil percobaan algoritma *Naïve Bayes* dan *Information Gain* mendapatkan hasil terbaik pada k = 55, yaitu akurasi sebesar 86,67%. *Precision* untuk kategori kritik adalah 84,6%, saran adalah 89,4% dan pertanyaan adalah 86,3%. *Recall* untuk kategori kritik adalah 88%, saran 84% dan pertanyaan 88%. Hal ini menyatakan bahwa sebagian besar klasifikasi dari aduan sudah sesuai. Kesalahan dalam klasifikasi juga tergantung pada data *training* yang kita ujikan. Penggunakan seleksi atribut *Information Gain* sangat berpengaruh pada hasil klasifikasi, karena jumlah atribut yang dipilih mempengaruhi nilai akurasinya. Banyak ataupun sedikit atribut yang digunakan mempengaruhi jumlah data yang diklasifikasi pada setiap kategori. Pada penelitian ini dengan menggunakan seleksi atribut kategori yang sering berubah ada pada kategori saran dan kritik. Hal ini disebabkan kata yang ada pada kategori kritik terdapat juga pada kategori saran sehingga pada menyebabkan kesalahan dalam pengklasifikasian di kedua kategori tersebut.

5.2. Saran

Berdasarkan kesimpulan yang telah dijelaskan di atas bahwa hasil klasifikasi pengaduan masyarakat dengan menggunakan algoritma *Naïve Bayes* dan *Information Gain* untuk seleksi atributnya meningkat. Namun tidak menutup kemungkinan bahwa masih terdapat kekurangan dalam proses pengujiannya. Untuk itu ada beberapa saran bagi penelitian-penelitian selanjutnya yang menggunakan algoritma yang sama dengan penelitian ini adalah sebagai berikut :

a. Untuk dapat meningkatkan tingkat akurasi dengan menggunakan algoritma *Naïve Bayes* dan seleksi atribut dengan menggunakan *Information Gain* sebaiknya data *training* dilakukan normalisasi secara manual dan secara detail.
b. Pada tahapan stopword dapat ditambahkan kata-kata yang dianggap tidak memberikan konstribusi pada setiap kategori aduan.

Pengujian seleksi atribut dapat dilakukan dengan menggunakan jumlah atribut yang diseleksi (k) dengan nilai yang lebih dekat, contohnya dimulai dari 5 sampai 100 dengan kelipatan 2 atau 3. Hal ini

untuk mengetahui nilai akurasi terbaik ada pada jumlah (k*)* yang ke berapa.

UCAPAN TERIMAKASIH

Penulis menyampaikan terima kasih kepada Rektor Universitas Dian Nuswantoro, Kaprodi MTI Universitas Dian Nuswantoro, dan semua pihak yang telah membantu penulis dalam penyusunan Tesis ini.

PERNYATAAN ORISINALITAS

" Saya menyatakan dan bertanggung jawab dengan sebenarnya bahwa artikel ini adalah hasil karya saya sendiri kecuali cuplikan dan ringkasan yang masing-masing telah saya jelaskan sumbernya"

[ALTER LASARUDIN]

DAFTAR PUSTAKA

[1] Adiwijaya, Igg. *Text Mining dan Knowledge Discovery*. Kolokium Bersama Komunitas Datamining Indonesia & Soft-Computing Indonesia, 2006.

[2] http://p2kp.org/pengaduandetil.asp?mid=37&catid=5&, 2016.

[3] Utami. L. D. *Integrasi Metode Information Gain Untuk Seleksi Fitur dan Adaboost Untuk Mengurangi Bias Pada Analisis Sentimen Review Restoran Menggunakan Algoritma Naïve Bayes*. Journal of Intelligent Systems, Vol. 1, No. 2, December 2015.

[4] Jaya Ivan. *Analisis Seleksi Atribut Pada Algoritma Naïve Bayes Dalam Memprediksi Penyakit Jantung*. Tesis Ilmu Komputer Dan Teknologi Informasi Universitas Sumatera Utara, 2013.

[5] Sulistyaningsih, Y, A. Djunaidy, Renny P. K. *Pengklasifikasian Pengaduan Masyarakat pada Laman Kantor Pertanahan Kota Surabaya I dengan Metode Pohon Keputusan*. Jurusan Sistem Informasi, Fakultas Teknologi Informasi Institut Teknologi Sepuluh Nopember. Surabaya, 2013.

[6] Darujati, Cahyo dan Agustinus B. Gumelar. *Pemanfaatan Teknik Supervised Untuk Klasifikasi Teks Bahasa Indonesia.* Jurnal Link Vol 16/No. 1/Februari 2012.

[7] N. S. Betha. *Implementasi Teknik Seleksi Fitur Information Gain Pada Algoritma Klasifikasi Machine Learning Untuk Prediksi Performa Akademik Siswa*. Seminar Nasional Teknologi Informasi dan Multimedia 2016. STMIK AMIKOM Yogyakarta, 6-7 Februari 2016.

[8] P. W. Akhmad dan Heru A. Santoso. *Naive Bayes Classification pada Klasifikasi Dokumen Untuk Identifikasi Konten E-Government*. Journal of Applied Intelligent System, Vol.1, No. 1, Februari 2016: 48-55.

[9] Amir Hamzah. *Klasifikasi Teks Dengan Naïve Bayes Classifier (Nbc) Untuk Pengelompokan Teks Berita Dan Abstract Akademis.* Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III. Yogyakarta, 3 Nopember 2012.

[10] Alves, Andre L. F, dkk. *A Comparison of SVM Versus Naive-Bayes Techniques for Sentiment Analysis in Tweets*: A Case Study with the 2013 FIFA Confederations Cup. WebMedia '14: Proceedings of the 20th Brazilian Symposium on Multimedia and the Web, November 2014.

[11] https://github.com/sastrawi/sastrawi/tree/master/src/Sastrawi/Stemmer. Diakses pada tanggal 20 Oktober 2016.