Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 2, No. 2, Februari 2018, hlm. 594-601 http://j-ptiik.ub.ac.id

Klasifikasi Teks Pengaduan Pada Sambat Online Menggunakan Metode N-Gram dan Neighbor Weighted K-Nearest Neighbor (NW-KNN)

Annisya Aprilia Prasanti[1], M. Ali Fauzi[2], M. Tanzil Furqon[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]annisyaap@gmail.com, [2]moch.ali.fauzi@gmail.com, [3]m.tanzil.furqon@gmail.com

Abstrak

SAMBAT Online merupakan salah satu bentuk nyata *E-Government* berupa aplikasi pengaduan berbasis website yang disediakan oleh Dinas Komunikasi dan Informatika Kota Malang (Diskominfo Malang). Suatu teks pengaduan yang masuk akan dikategorikan ke dalam berbagai bidang SKPD yang bertanggung jawab. Untuk mempermudah mengorganisir teks pengaduan dan meningkatkan efisiensi waktu *super admin* dalam memilah dan menentukan bidang SKPD tujuan, diperlukan sebuah metode klasifikasi teks. NW-KNN merupakan algoritme pengembangan dari algortime KNN tradisional. Umumnya, perhitungan jarak tetangga terdekat yang digunakan pada algoritme NW-KNN menggunakan *Cosine Similarity* dengan ekstraksi fitur *bag of words.* B*ag of words* merupakan ekstraksi fitur yang tidak memperhatikan urutan dari kata sebuah kalimat. Untuk menyempurnakan kekurangan tersebut, pada penelitian ini akan digunakan metode pendukung untuk ekstrasi fitur yaitu metode N-Gram. Hasil pengujian dalam penelitian ini menunjukkan bahwa penggunaan metode NW-KNN dengan nilai tetangga k = 3 dan metode N-Gram dengan Unigram memiliki nilai *f-measure* tertinggi sebesar 75.25%.

Kata kunci: *Klasifikasi Teks, Text Mining, Neighbor Weighted K-Nearest Neighbor, N-Gram*

Abstract

SAMBAT Online is a concrete application of E-Government in a web-based platform for complaints provided by Dinas Komunikasi dan Informatika Kota Malang (Diskominfo Malang). An incoming complaint text will be categorized into various areas of the SKPD. With that being said, in order to make the job of the super admin easier in organizing and determining an SKPD category, as well to organize a complaint text and improve the time efficacy, a method of text classification is paramount. NW-KNN is an upgraded algorithm of the traditional KNN algorithm. Generally, the closest neighboring distance calculations will use Cosine Similarity with bag of words for feature extraction. Bag of words is a feature extraction that ignores the order of words of a sentence altogether. To improve the algorithm despite the deficiency, this research will use supporting method for feature extraction, which is called as N-Gram. The result in this research indicated that NW-KNN with neighboring value k = 3 and N-Gram with Unigram have the highest f-measure's value with 75.25%.

Keywords: *Text Classification, Text Mining, Neighbor Weighted K-Nearest Neighbor, N-Gram*

1. PENDAHULUAN

Era globalisasi secara tidak langsung memaksa masyarakat untuk dapat menyesuaikan diri agar tidak menjadi pihak terbelakang. Salah satu bentuk nyatanya adalah pemanfaatan teknologi informasi dan komunikasi (ICT) dalam wujud *E-Government*. Dengan adanya *E-Government*, diharapkan komunikasi dua arah dapat dilakukan oleh masyarakat kepada pemerintah guna menyampaikan aspirasi mereka untuk menilai, mengkritik, bertanya atau

memberikan opini mereka (Anandita, 2016).

SAMBAT Online (Sistem Aplikasi Masyarakat Bertanya Terpadu Online) merupakan salah satu pemanfaatan *E-Government* yang disediakan oleh Diskominfo (Dinas Komunikasi dan Informatika) Kota Malang. SAMBAT Online merupakan aplikasi yang ditujukan untuk masyarakat kota Malang guna mengirimkan saran, kritik, pertanyaan atau pengaduan seputar penyelenggaraan pelayanan atau fasilitas publik oleh pemerintah Kota Malang.

Dalam penerapan aplikasi SAMBAT Online, Diskominfo Kota Malang bertugas sebagai super admin untuk menerima seluruh pengaduan yang masuk. Selanjutnya, super admin harus memilah dan meneruskan (*forward*) pengaduan ke SKPD yang bertanggung jawab atas perihal pengaduan tersebut secara manual. Hal ini kurang efisien karena jika terdapat pengaduan yang bersifat penting dan mendesak, maka tidak dapat ditangani secara cepat karena banyaknya jumlah pengaduan yang masuk serta bercampur dengan perihal pengaduan lain.

Untuk meningkatkan efisiensi waktu *super admin* dalam memilah dan menentukan SKPD tujuan, diperlukan sebuah metode klasifikasi teks. Berbagai metode telah diterapkan untuk menyelesaikan masalah klasifikasi teks, salah satu diantaranya adalah metode *Neighbor Weighted K-Nearest Neighbor* (NW-KNN). Metode NW-KNN merupakan metode pengembangan dari metode *K-Nearest Neighbor* (KNN) tradisional dengan menambahkan tahapan pembobotan untuk menyelesaikan permasalahan *corpus* yang tidak seimbang (Sangbo, 2005). Algoritme NW-KNN memberikan nilai bobot kecil untuk tetangga yang berasal dari kelas mayoritas dan memberikan nilai bobot besar untuk tetangga yang berasal dari kelas minoritas. Penelitian Sangbo Tan (2005) memperoleh hasil peningkatan yang signifikan pada persebaran data tidak seimbang. Hal tersebut terbukti pada hasil penelitiannya yang memaparkan bahwa performa metode NW-KNN lebih baik dibanding metode KNN tradisional.

Umumnya, perhitungan jarak tetangga terdekat pada algoritme KNN adalah dengan perhitungan *Cosine Similarity* menggunakan *bag of words* (Utomo, 2015). B*ag of words* merupakan ekstraksi fitur yang tidak memperhatikan urutan kata sebuah kalimat. Jika terdapat lebih dari satu kalimat yang berbeda dengan komposisi kata yang sama, maka kalimat tersebut akan dianggap mirip dan hal tersebut akan mempengaruhi akurasi. Untuk menyempurnakan kekurangan yang dimiliki, maka penelitian ini akan digunakan metode pendukung untuk ekstrasi fitur yaitu N-Gram.

Djoko Cahyo Utomo (2015) pada penelitiannya yang berjudul *Automatic Essay Scoring (AES)* menggunakan metode *N-gram* dan *Cosine similarity* memaparkan perbedaan hasil korelasi pada 2 *dataset* yang digunakan. Pada *dataset* 1, nilai Unigram lebih bagus nilai korelasinya antara 0,63-0,64. Sedangkan pada *dataset* 2, nilai Bigram dan nilai kombinasi lebih bagus. Nilai korelasi Bigram antara 0,6-0,66 dan nilai korelasi kombinasi antara 0,62-0,67.

Selain itu, penelitian lain yang dilakukan oleh Yudha Permadi (2008) berjudul Kategorisasi Teks Menggunakan N-Gram untuk Dokumen Berbahasa Indonesia memaparkan hasil bahwa N-Gram Trigram mampu mengklasifikasikan dokumen dengan benar dengan presentase tertinggi sebesar 81,25% dan presentase terendah sebesar 5,357%.

Dengan menerapkan metode NW-KNN yang didukung oleh ekstrasi fitur N-Gram, diharapkan performa sistem klasifikasi teks pengaduan pada SAMBAT Online menggunakan metode N-Gram dan NW-KNN mampu menyelesaikan permasalahan klasifikasi teks serta menjadi acuan untuk membangun model klasifikasi teks pengaduan pada sistem SAMBAT Online.

2. KAJIAN PUSTAKA

2.1. SAMBAT Online

SAMBAT Online merupakan salah satu bentuk nyata dari aplikasi pengaduan berbasis website yang disediakan oleh Dinas Komunikasi dan Informatika Kota Malang. Sistem Aplikasi Masyarakat Bertanya Terpadu Online atau yang

disingkat dengan SAMBAT Online merupakan aplikasi yang ditujukan untuk masyarakat kota Malang guna mengirimkan saran, kritik, pertanyaan atau pengaduan seputar penyelenggaraan pelayanan atau fasilitas publik oleh pemerintah Kota Malang.

2.2. *Text Mining*

Text mining dapat didefinisikan sebagai penambangan data yang digunakan untuk mendapat kata atau pola sebagai wakil isi kumpulan teks atau dokumen sehingga dapat diproses dengan tujuan tertentu. *Text mining* secara umum bertujuan untuk mendapatkan informasi penting dari sekumpulan teks atau dokumen (Witten, 2003).

Berdasarkan sifat teks yang tidak terstruktur dan kompleks, maka proses *text mining* memerlukan beberapa tahap yang pada intinya untuk mempersiapkan teks agar dapat diubah menjadu lebih terstruktur. Tahap awal yang dilakukan adalah dengan *preprocessing* yang dimana *text mining* tidak hanya melakukan perhitungan dokumen, namun juga perhitungan beberapa fitur yaitu *character*, *words*, *terms*, dan *concept* (Feldman & Sanger, 2007).

2.3. Klasifikasi

Klasifikasi merupakan tahap menemukan suatu model yang menggambarkan atau membedakan kelas data yang bertujuan untuk memprediksi kelas dari suatu objek baru (Han & Kamber, 2006).

Telah berkembang berbagai metode yang dapat diterapkan untuk permasalahan klasifikasi teks, salah satunya adalah metode NW-KNN. Tahapan yang dilakukan untuk melakukan klasifikasi teks antara lain sebagai berikut.

1. *Preprocessing*

Preprocessing merupakan sebuah proses yang bertujuan untuk mempersiapkan dokumen mentah sebelum diolah, baik dari dokumen latih maupun dokumen uji. Fungsi *preprocessing* data diimplementasikan untuk memindahkan dokumen awal ke dalam representasi yang rapi serta melakukan implementasi pada dokumen uji maupun dokumen latih (Guo, Wang, Bell, Bi, & Greer, 2004). Tahapan-tahapan yang terdapat dalam *text preprocessing* antara lain *tokenizing*, *filtering/stopword removal*, dan *stemming*.

2. *Term Weighting*

Term weighting (pembobotan) merupakan proses merubah data kualitatif menjadi data kuantitatif sehingga dapat diproses dengan perhitungan komputasi pada komputer. Setelah data telah selesai melalui tahap preprocessing, maka selanjutnya data tersebut masuk ke tahap pembobotan. Metode yang paling banyak digunakan untuk melakukan pembobotan adalah *Term Frequency* (TF) dan *Term Frequency-Inverse Document Frequency* (TF-IDF).

3. Klasifikasi Teks

Teknik klasifikasi digunakan untuk melakukan prediksi atas informasi yang belum diketahui sebelumnya (Suprawoto, 2016).

Terdapat 4 komponen utama yang menjadi dasar proses klasifikasi (Gorunescu, 2011), antara lain sebagai berikut.

a) Kelas: Merupakan variabel terikat (dependen) yang berupa kategorikal untuk merepresentasikan suatu label pada suatu objek.
b) *Predictor*: Merupakan variabel independen yang merepresentasikan karakteristik suatu data.
c) *Training Dataset*: Merupakan suatu kumpulan data yang berisi nilai dari komponen kelas dan *predictor* yang akan berguna untuk penentuan kelas data uji berdasarkan *predictor*.
d) *Testing Dataset*: Merupakan suatu kumpulan data yang akan diklasifikasikan oleh model yang telah terbentuk.

2.4. N-Gram

N-Gram merupakan potongan *n*-karakter yang diperoleh dari sebuah *string*. Metode N-Gram diaplikasikan untuk pembangkitan kata atau karakter. Pada umumnya, N-Gram yang utuh didapat dengan menambahkan *blank* di awal dan di akhir suatu kata. Misalnya, terdapat kata "TEKS" yang dapat diuraikan ke dalam beberapa n-gram berikut yang dimana "_" merepresentasikan *blank* (Utomo, 2015).

Unigram : T,E,K,S
Bigram : _T, TE, EK,KS, dan S

Trigram : _TE,TEK,EKS, KS_ dan S_ _

Pada penelitian ini, N-Gram akan diaplikasikan dengan modifikasi pemecahan/pemisahan berdasarkan per-kata, tidak per-huruf *n*-karakter. Seperti contohnya, terdapat kalimat berikut; 'klasifikasi teks pengaduan pada sambat online menggunakan metode N-Gram dan NW-KNN'

Unigram : klasifikasi, teks, pengaduan, pada, sambat, online, menggunakan, metode, N-Gram, dan, NW-KNN.

Bigram : klasifikasi teks, teks pengaduan, pengaduan pada, pada sambat, sambat online, online menggunakan, menggunakan metode, metode N-Gram, N-Gram dan, dan NW-KNN.

Gabungan : klasifikasi, teks, pengaduan, pada, sambat, online, menggunakan, metode, N-Gram, dan, NW-KNN, klasifikasi teks, teks pengaduan, pengaduan pada, pada sambat, sambat online, online menggunakan, menggunakan metode, metode N-Gram, N-Gram dan, dan NW-KNN.

2.5. *Neighbor Weighted K-Nearest Neighbor* (NW-KNN)

Dugaan bahwa data latih yang tersimpan didistribusikan secara menyeluruh dan seimbang pada tiap kelas tidak selalu tepat. Pada data yang tidak seimbang, kelas mayoritas direpresentasikan dengan banyak presentase data latih sedangkan kelas minoritas hanya memiliki sedikit presentase data latih.

Sangbo Tan (2005) mengemukakan algoritme *Neighbor Weighted K-Nearest Neighbor* (NW-KNN) untuk data tidak seimbang pada struktur kumpulan teks, yang dimana pada penelitiannya mengungkapkan bahwa performa klasifikasi teks menurun ketika algoritme KNN tradisional menghadapi data tidak seimbang.

Penerapan algoritme NW-KNN tidak jauh berbeda dengan algoritme KNN tradisional. Tahap awal adalah dengan melakukan perhitungan jarak (kemiripan) antara objek data uji dengan kelompok *k* data latih. Setelah mendapat nilai kemiripan (similaritas), selanjutnya dilakukan pengurutan nilai kemiripan (similaritas) tersebut berdasarkan nilai terbesar dari pemilihan *k* tetangga. Perbedaan keduanya terletak pada perhitungan skor. Pada algoritma NW-KNN, akan dilakukan perhitungan bobot sebelum perhitungan skor. Pada perhitungan ini, kategori minoritas akan diberi bobot lebih besar, sedangkan kategori mayoritas akan diberi bobot lebih kecil. Rumus perhitungan bobot dihitung dengan persamaan berikut (Sangbo, 2005).

$$Weight_i = \frac{1}{\left(\frac{Num\left(C_i^d\right)}{Min\{Num(C_j^d) | j = 1, \dots, k*\}}\right)^{1/exp}} \quad (1)$$

Keterangan :

$Num\left(C_i^d\right)$ = banyaknya data latih *d* pada kategori *i*

$Min\{Num(C_j^d) | j = 1, \dots, k*\}$ = banyaknya data latih minimum dari seluruh kategori

exp = eksponen, bilangan lebih dari 1

Setiap bobot kategori yang didapatkan kemudian akan digunakan untuk menghitung skor data uji q terhadap setiap kategori. Hasil dari penghitungan skor akan digunakan sebagai acuan untuk menentukan kategori dari data uji yang diproses. Perhitungan skor setiap data uji untuk metode NWKNN dapat dihitung dengan Persamaan (2).

$$Score(q, Ci) = Weight_i \left(\sum_{dj \in KNN(q)} Sim(q, dj)\ \delta(dj, Ci) \right) \quad (2)$$

Keterangan :

$Weight_i$ = bobot kelas

$dj \in KNN(q)$ = data latih d_j yang berada pada kumpulan tetangga terdekat (nearest neighbor) dari dokumen uji q

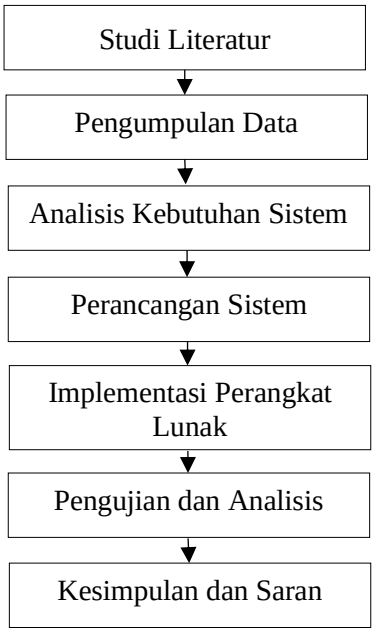
$Sim(q, dj)$ = similaritas antara dokumen uji q dengan dokumen latih d_j

$\delta(dj, Ci)$ = $\begin{cases} dj \in Ci = 1 \\ dj \notin Ci = 0 \end{cases}$

Ci = kelas / kategori i

3. METODOLOGI PENELITIAN

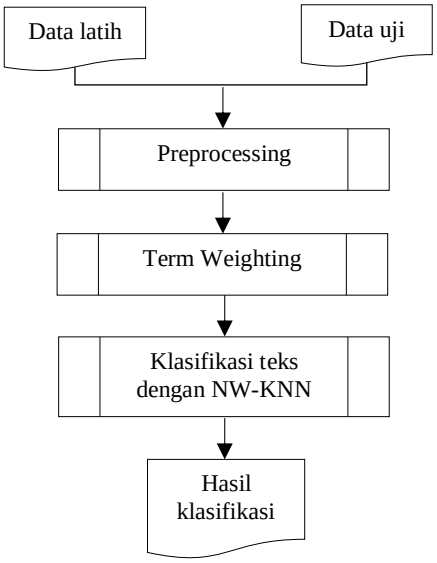
Tahapan penelitian yang dilakukan meliputi studi literatur, pengumpulan data, analisis kebutuhan sistem, perancangan sistem, implementasi perangkat lunak, pengujian dan analisis serta kesimpulan dan saran.



Gambar 1. Alur Metode Penelitian

3.1. Alur Kerja Sistem

Pada penelitian ini, terdapat tiga proses utama yang akan dilakukan oleh sistem yaitu proses *preprocessing*, proses *term weighting* dan proses klasifikasi dengan metode NW-KNN. Secara garis besar, alur kerja sistem klasifikasi teks pengaduan pada SAMBAT Online menggunakan metode N-Gram dan NW-KNN direpresentasikan pada Gambar 1. Tahapan alur kerja sistem terdiri dari beberapa bagian utama, antara lain sebagai berikut.



Gambar 2. Alur Kerja Sistem

a) Preprocessing

Pada tahap *preprocessing*, akan dilakukan beberapa proses yang bertujuan untuk memecah teks pengaduan menjadi beberapa term agar siap diolah pada tahap berikutnya. Tahap *preprocessing* ini meliputi beberapa proses antara lain *tokenizing*, *filtering* dan *stemming*. Setelah melalui proses ini, sistem akan melakukan proses ekstrasi fitur dengan menggunakan N-Gram.

b) Term Weighting

Perhitungan pembobotan yang dilakukan pada penelitian ini menggunakan metode pembobotan TF-IDF. Nilai TF merupakan jumlah kemunculan jumlah kata pada suatu dokumen. Sedangkan IDF merupakan pembobotan kata yang didasarkan pada banyaknya dokumen yang mengandung kata tertentu.

c) Klasifikasi teks dengan NW-KNN

Klasifikasi teks merupakan proses terakhir pada penelitian ini. Pada proses ini, terdapat beberapa tahap yang harus dilalui yaitu tahap perhitungan similaritas dokumen dengan metode *cosine similarity* (CosSim), tahap perhitungan skor akhir dan tahap pemilihan maksimum skor. Proses ini akan menghasilkan nilai terbesar skor yang mewakilkan kategori dari klasifikasi dokumen dengan algoritme NW-KNN.

4. HASIL DAN PEMBAHASAN

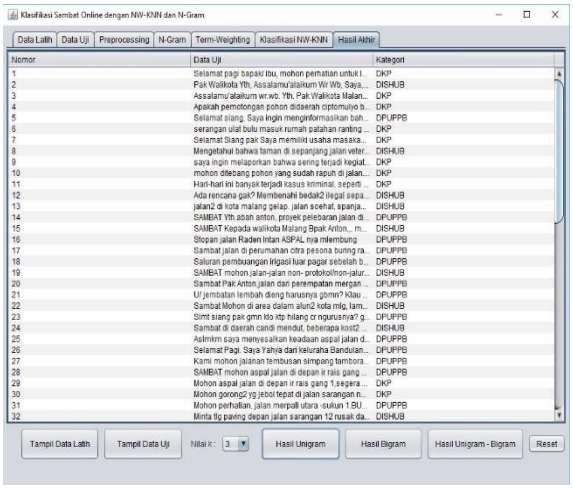
Data yang digunakan pada penelitian ini diambil dari wesite SAMBAT Online di *http://sambat.malangkota.go.id*.

Teks pengaduan diambil dari 3 SKPD antara lain Dinas Perhubungan (DISHUB), dan Dinas Kebersihan dan Pertamanan (DKP), dan Dinas Pekerjaan Umum, Perumahan dan Pengawasan Bangunan (DPUPPB). Total seluruh data berjumlah 310 data dengan uraian 237 data latih dan 73 data uji. Data latih yang digunakan pada kelas DKP berjumlah 27 data, pada kelas DPUPPB berjumlah 49 data dan pada kelas DISHUB berjumlah 161 data. Sedangkan data uji yang digunakan pada kelas DKP berjumlah 13 data, pada kelas DPUPPB berjumlah 21 data dan pada kelas DISHUB berjumlah 39 data.

Implementasi sistem pada penelitian ini menggunakan bahasa pemrograman Java berbasis *desktop*. Menu-menu yang tersedia pada sistem antara lain Data Latih, Input Data Uji, Preprocessing, N-Gram, Term Weighting, Klasifikasi NW-KNN, dan Hasil Akhir. Gambar 3 menunjukkan implementasi antarmuka Hasil Akhir. Contoh hasil perhitungan skor akhir dari klasifikasi NW-KNN oleh sistem ditunjukkan pada Tabel 1.

Tabel 1. Contoh Hasil Perhitungan Skor

Dok. Uji	Nilai Perhitungan Skor			Hasil Akhir
	DISHUB	DPUPPB	DKP	
D_1	0.0	0.184	0.404	DKP
D_2	0.290	0.0	0.0	DISHUB
…	…	…	…	…
D_{73}	0.178	0.0	0.0	DISHUB



Gambar 3. Implementasi Sistem

Pada penelitian ini, dilakukan pengujian hasil klasifikasi sistem untuk dianalisis. Pengujian yang dilakukan adalah menguji nilai performa sistem diukur dengan menggunakan konsep *precision, recall* dan *f-measure*.

4.1. Pengaruh Variasi Nilai *k*

Pada pengujian ini, dilakukan perbandingan nilai *k* dari 5 variasi nilai *k* yaitu 1, 3, 5, 7 dan 15. Tabel 2 berikut menunjukkan hasil rata-rata perhitungan *precision, recall* dan *f-measure* dari tiap nilai *k* dari seluruh kelas.

Tabel 2 Hasil Pengujian Pengaruh Variasi Nilai k

Nilai k	Precision	Recall	F-Measure
1	69.60%	63.51%	65.51%
3	**77.85%**	**74.18%**	**75.25%**
5	75.13%	68.31%	70.60%
7	76.51%	68.31%	70.95%
15	74.02%	64.50%	67.02%

Berdasarkan Tabel 2, nilai *k* = 3 memiliki hasil klasifikasi paling optimal. Hal ini disebabkan karena jika jumlah tetangga data latih yang dipertimbangkan sangat sedikit, maka data uji tidak memiliki probabilitas lain untuk masuk ke kelas lainnya. Hasil pengujian ini juga menunjukkan penurunan performa sistem ketika nilai *k* yang digunakan semakin besar. Hal ini disebabkan karena semakin besar nilai *k*, maka semakin besar pula probabilitas tetangga yang jaraknya jauh ikut dipertimbangkan. Hal ini bertolak belakang dengan prinsip kerja algoritme

KNN yang melakukan klasifikasi dengan mencari jarak yang paling dekat terhadap kelompok *k* data latih.

4.2. Pengaruh Variasi N-Gram

Pada pengujian ini, variasi N-Gram yang digunakan adalah Unigram, Bigram dan Unigram-Bigram. Tabel 3 berikut menunjukkan hasil rata-rata perhitungan *precision*, *recall* dan *f-measure* dengan menggunakan nilai *k* = 3.

Tabel 3 Hasil Pengujian Variasi N-Gram

N-Gram	Precision	Recall	F-Measure
Unigram	**77.85%**	**74.18%**	**75.25%**
Bigram	55.85%	46.44%	48.51%
Unigram-Bigram	70.51%	69.57%	69.57%

Berdasarkan Tabel 3, fitur N-Gram dengan Unigram menunjukkan hasil rata-rata nilai *f-measure* paling optimal dibanding dua fitur N-Gram lainnya. Hal ini disebabkan karena banyak *term* Bigram yang jarang muncul pada lebih dari satu dokumen. Fitur Bigram merupakan penggabungan dua buah kata yang dijadikan sebagai satu *term.* Pada penelitian ini, mayoritas satu *term* Bigram yang muncul hanya ada pada satu dokumen, yaitu dokumen dimana *term* itu berada.

4.3. Pengaruh Algoritme KNN dan Algoritme NW-KNN

Pada pengujian ini, dilakukan perbandingan antara algoritme KNN dengan algoritme NW-KNN. Variasi nilai tetangga *k* yang digunakan antara lain 1, 3, 5, 7, 15 dan N-Gram yang digunakan adalah Unigram. Tabel 4 berikut menunjukkan hasil rata-rata perhitungan *precision*, *recall* dan *f-measure* dengan menggunakan algoritme KNN dengan algoritme NW-KNN.

Tabel 4. Hasil Pengujian Pengaruh Algoritme KNN dan Algoritme NW-KNN

Algoritme KNN			
Nilai k	**Precision**	**Recall**	**F-Measure**
1	69.60%	63.51%	65.51%
3	85.91%	70.52%	75.21%
5	81.51%	66.72%	70.83%
7	86.63%	60.03%	62.51%
15	85.27%	60.03%	62.15%
Algoritme NW-KNN			
Nilai k	**Precision**	**Recall**	**F-Measure**
1	69.60%	63.51%	65.51%
3	**77.85%**	**74.18%**	**75.25%**
5	75.13%	68.31%	70.60%
7	76.51%	68.31%	70.95%
15	74.02%	64.50%	67.02%

Berdasarkan Tabel 4, algoritme NW-KNN menunjukkan hasil rata-rata *f-measure* lebih baik dibanding algoritme KNN biasa ketika nilai tetangga *k* bernilai besar. Hal ini disebabkan karena semakin besar nilai tetangga *k,* maka semakin besar pula probabilitas kelas data latih yang sesungguhnya bukan tetangga terdekat yang ikut dipertimbangkan. Hal tersebut juga berpengaruh karena sebaran jumlah data latih pada tiap kelas tidak seimbang. Saat nilai tetangga *k* semakin besar dengan menggunakan algoritme KNN, suatu data uji akan memiliki probabilitas kesalahan untuk masuk ke kelas yang memiliki jumlah data latih terbanyak. Sedangkan pada algoritme NW-KNN, terdapat perhitungan bobot kelas pada yang mampu meminimalisir probabilitas kesalahan tersebut sehingga hasil klasifikasi menjadi lebih akurat.

5. KESIMPULAN

Berdasarkan hasil pengujian penelitian ini, dapat disimpulkan bahwa algoritme NW-KNN mampu melakukan klasifikasi teks pengaduan dengan nilai tetangga *k* terdekat yang paling optimal adalah 3, dengan rata-rata hasil presentase *percision* sebesar 77.85%, rata-rata hasil presentase *recall* sebesar 74.18% dan rata-rata hasil presentase *f-measure* sebesar 75.25%.

Penelitian ini juga menunjukkan bahwa metode N-Gram yang diterapkan pada sistem tidak memiliki pengaruh yang besar terhadap hasil klasifikasi. Pernyataan ini didukung hasil pengujian variasi N-Gram yang menyatakan

bahwa Unigram memiliki nilai rata-rata *f-measure* paling besar dengan presentase 75.25% dibanding Bigram dan Unigram-Bigram.

Pada penelitian ini, penggunaan algoritme NW-KNN terbukti lebih baik untuk permasalahan data tidak seimbang ketika nilai tetangga *k* bernilai besar daripada penggunaan algoritme KNN biasa. Hal ini disebabkan karena saat nilai tetangga *k* bernilai besar, suatu data uji akan memiliki probabilitas kesalahan untuk masuk ke kelas yang memiliki jumlah data latih terbanyak jika menggunakan algoritme KNN biasa.

Dari hasil pengujian, analisis, dan kesimpulan yang telah dirumuskan, terdapat beberapa hal yang dapat dikembangkan sebagai bahan untuk penelitian selanjutnya, antara lain perlu adanya proses langsung teks pengaduan dari sistem SAMBAT Online dalam bentuk **.html* atau sejenisnya sehingga tidak perlu memindahkan teks pengaduan ke dalam file **.xls*. Selain itu, penelitian ini perlu adanya tambahan proses untuk mendeteksi kata yang disingkat dan kata tidak baku karena mayoritas pengguna sistem SAMBAT Online merupakan masyarakat umum yang sering menggunakan kata tidak baku.

DAFTAR PUSTAKA

Anandita, N. 2016. *Elemen Sukses E – Government: Studi Kasus Layanan Aspirasi Dan Pengaduan Online Rakyat (Lapor!) Kota Bandung.* Universitas Katolik Parahyangan, Bandung.

Feldman, R. & Sanger, J., 2007. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, New York.

Guo, G. et al., 2004. *An KNN Model-Based Approach and Its Application in Text Categorization.* Springer-Verlag Berlin Heidelberg, UK.

Gorunescu, F., 2011. *Data Mining: Concepts, Models, and Techniques.* Springer, Verlag Berlin Heidelberg.

Han, J. & Kamber, M., 2006. *Data Mining Concept and Tehniques.* Morgan Kauffman, San Fransisco.

Indriati & Ridok, A., 2016. *Sentiment Analysis For Review Mobile Applications Using Neighbor Method Weighted K-Nearest Neighbor (Nwknn).* Universitas Brawijaya, Malang.

Permadi, Y., 2008. *Kategorisasi Teks Menggunakan N-Gram Untuk Dokumen Berbahasa Indonesia.* Bogor: Departemen Ilmu Komputer, FMIPA IPB.

Suprawoto, T., 2016. *Klasifikasi Data Mahasiswa Menggunakan Metode K-Means Untuk Menunjang Pemilihan Strategi Pemasaran.* STMIK AKAKOM, Yogyakarta.

Tan, S. 2005. *Neighbor-weighted K-nearest neighbor for unbalanced text corpus. Expert Systems with Applications 28 (2005) 667–671.*

Utomo, D. C., 2015. *Automatic Essay Scoring (AES) Menggunakan Metode N-Gram dan Cosine Similarity.* Universitas Brawijaya, Malang.

Witten, I. H., 2003. *Text Mining.* University of Waikato, New Zealand.