

## **Perbandingan Pengklasifikasi *k*-Nearest Neighbor dan Neighbor-Weighted *k*-Nearest Neighbor Pada Sistem Analisis Sentimen dengan Data Microblog**

**Martha<sup>1</sup>**

Universitas Tarumanagara  
e-mail: [viny@untar.ac.id](mailto:viny@untar.ac.id)

**Viny Christanti M<sup>2</sup>**

Universitas Tarumanagara

**Dali S. Naga<sup>3</sup>**

Universitas Tarumanagara

**P. T. D. Rompas<sup>4</sup>**

Universitas Negeri Manado

### **ABSTRAK**

Sistem analisis sentimen merupakan suatu sistem yang dibangun untuk menganalisis sentimen dalam bahasa Indonesia pada data yang diambil dari microblog, yaitu Twitter, Facebook, dan YouTube. Data tersebut diklasifikasikan menggunakan K-Nearest Neighbor (KNN) dan Neighbor-Weighted K-Nearest Neighbor (NW-KNN). Sistem analisis sentimen dibagi menjadi 4 (empat) tahap, yaitu tahap microblog crawling, tahap pra-pemrosesan data, tahap ekstraksi fitur, dan tahap klasifikasi sentimen. Data yang ada akan diklasifikasikan ke dalam 5 (lima) kelas, yaitu sangat positif, positif, netral, negatif, dan sangat negatif. Data tersebut kemudian akan dievaluasi menggunakan confusion matrix. Pengujian dilakukan dengan 50 data uji. NW-KNN memiliki akurasi yang paling tinggi di antara seluruh pengklasifikasi, yaitu 86% pada saat  $K = 9$ . Sedangkan untuk KNN, akurasi dari pengklasifikasi tersebut sebesar 82% pada saat  $K = 3$ . Kesimpulannya adalah NW-KNN berhasil mengatasi data latih dengan komposisi kelas yang tidak seimbang.

***Kata kunci: KNN, NW-KNN, Analisis Sentimen, Data Microblog***

### **ABSTRACT**

*Sentiment analysis system is a system built to analyze sentiments in Indonesian on data taken from microblog, ie Twitter, Facebook, and YouTube. The data is classified using K-Nearest Neighbor (KNN) and Neighbor-Weighted K-Nearest Neighbor (NW-KNN). Sentiment analysis system is divided into 4 (four) stages, namely microblog crawling stage, pre-processing data stage, feature extraction stage, and classification phase of sentiment. Existing data will be classified into 5 (five) classes, which are very positive, positive, neutral, negative, and very negative. The data will then be evaluated using a confusion matrix. Testing was done with 50 test data. NW-KNN has the highest accuracy among all classifiers, ie 86% at  $K = 9$ . As for KNN, the accuracy of the classifier is 82% at  $K = 3$ . In conclusion, NW-KNN successfully overcame the train data with composition unbalanced classes.*

***Keywords: KNN, NW-KNN, Sentimen Analysis, MicrobloG Data***

## PENDAHULUAN

Pada zaman yang terus-menerus berkembang ini, sentimen atau opini masyarakat semakin bertambah luas dan bebas diungkapkan di berbagai media. Menurut Kamus Besar Bahasa Indonesia (KBBI), sentimen adalah pendapat atau pandangan yang didasarkan pada perasaan yang berlebih-lebihan terhadap sesuatu (bertentangan dengan pertimbangan pikiran). Sentimen terdapat dalam berbagai macam topik yang ada, misalnya produk, layanan, pandangan, agama, politik, dan topik-topik lainnya. Sentimen dapat menjadi potensi besar bagi perusahaan yang ingin mengetahui umpan balik (*feedback*) dari masyarakat terhadap merk dagang mereka.

Sistem analisis sentimen (*sentiment analysis*) merupakan sistem yang dapat menganalisis sentimen yang ada dari suatu pernyataan. Sentimen dalam sistem dibagi menjadi 5 (lima) kelas, yaitu sangat positif, positif, netral, negatif, dan sangat negatif. Sentimen atau opini masyarakat dapat diperoleh dari kegiatan berkomunikasi, baik secara tertulis, maupun lisan. Salah satu kegiatan berkomunikasi secara tertulis adalah *microblogging*. Contoh dari kalimat pada setiap kelas sentimen ada pada tabel 1.

Tabel 1. Contoh Kalimat pada Setiap Kelas

KALIMAT	KELAS	KODE
baguus itu banget!	Sangat Positif	5
Mantap menambah wawasan	Positif	4
Antara bagus dan gk bagus.	Netral	3
Aku sedih nontonnya.	Negatif	2
liat muka sony koq eneq bgt y	Sangat Negatif	1

*Microblogging* adalah kegiatan untuk menyebarkan potongan-potongan kecil dari konten digital di internet, misalnya teks, gambar, tautan, video berdurasi pendek, atau media lainnya. Beberapa *microblog* yang digunakan adalah Facebook, Twitter, dan YouTube. Sistem analisis sentimen yang dibangun menggunakan 2 (dua) jenis algoritma klasifikasi teks berbasis statistik, yaitu *Neighbor-Weighted K-Nearest Neighbor* (NW-KNN) sebagai algoritma utama dan *K-Nearest Neighbor* (KNN) sebagai algoritma pembanding tingkat kinerja dan akurasi dari masing-masing algoritma klasifikasi.

Percobaan sebelumnya yang pernah dilakukan oleh Harlili dan Yudi Wibisono menunjukkan bahwa sistem analisis sentimen yang dibangun menggunakan 33000 tweet dengan metode *Naïve Bayes* dapat menghasilkan akurasi sejumlah 91,83%. Percobaan yang dicoba membuktikan pembobotan kelas pada NW-KNN dapat menangani masalah data yang tidak seimbang yang sering terdapat pada data *microblog*. Percobaan ini diharapkan dapat membantu pengembangan sistem analisis sentimen bahasa Indonesia. Selain itu, percobaan ini diharapkan dapat membantu masyarakat mengetahui sentimen terhadap suatu topik seperti produk, layanan, pandangan, agama, politik, dan topik-topik lainnya menurut pandangan para pengguna situs *microblog* dalam bahasa Indonesia.

## METODE

Penelitian ini berfokus pada pengklasifikasian sentimen pada koleksi data bahasa Indonesia yang diperoleh dari data *microblog* dengan membandingkan 2 (dua) algoritma klasifikasi, yaitu KNN dan NW-KNN. Secara garis besar, sistem analisis sentimen yang dibangun terdapat 4 (empat) proses utama, yaitu tahap

*microblog crawling*, tahap pra-pemrosesan data, tahap ekstraksi fitur, dan tahap klasifikasi sentimen.

Sistem analisis sentimen didukung oleh koleksi data yang didapat dari data pada *microblog*, seperti Twitter, Facebook, dan YouTube. Sistem menggunakan koleksi data berupa data latih dan kamus. Tabel 2 menunjukkan daftar jumlah kata dalam kamus yang digunakan pada saat pra-pemrosesan data.

Tabel 2. Daftar Jumlah Kata dalam Kamus

NO	NAMA KAMUS	JUMLAH KATA
1	abbreviations.dict	212 kata
2	words.dict	43214 kata
3	excessives.dict	50 kata
4	rejections.dict	37 kata
5	stopwords.dict	753 kata

“*abbreviations.dict*” merupakan kamus dengan daftar kata singkatan dan digunakan untuk mengubah kata singkatan pada data menjadi kata baku dari singkatan tersebut. “*words.dict*” digunakan untuk mengecek apakah kata dengan karakter berulang pada data bahasa Indonesia terdapat pada kamus. Apabila tidak terdaftar maka karakter berulang akan dihapus. “*excessives.dict*” merupakan kamus dengan daftar kata penguat dan digunakan pada saat proses penglabelan kata. Contoh kata penguat adalah “sangat”, “luar biasa”, “amat”, dan lain-lain. “*rejections.dict*” merupakan kamus dengan daftar kata negasi dan digunakan pada saat penglabelan, misalnya “tidak”, “gak”, “bukan”, dan lain-lain. “*stopwords.dict*” merupakan kamus berisi daftar *stopwords* dalam bahasa Indonesia.

Skema sistem analisis sentimen yang diteliti dapat dilihat pada Gambar 1. Pada tahap awal dilakukan pencarian data dengan *web crawling* dan mencari data

pada *microblog*. Pada tahap *microblog crawling* atau *web crawling*, pencarian data dari setiap situs *microblog* yang dipilih dilakukan sesuai dengan kata kunci yang dimasukkan oleh pengguna. *Crawling* dilakukan terhadap 3 (tiga) situs *microblog* utama, yaitu Twitter, Facebook, dan YouTube. Data yang telah di-*crawl* dibersihkan dengan membuang *username*, *hashtag*, *retweet*, dan tautan Uniform Resource Locator (URL) yang biasanya tidak mengandung sentimen. Setelah proses *crawling* selesai, data hasil *crawling* akan disimpan dan diberi kelas secara manual.

Langkah selanjutnya adalah pra-pemrosesan data yang ada, yaitu data latih dan data uji, yang diawali dengan tokenisasi (*tokenizing*), pengubahan huruf besar menjadi huruf kecil (*case folding*), pengecekan kata (*word checking*), penjernihan *noise* (*cleaning*), penglabelan tanda seru, tanda tanya, *negation* dan *exaggeration* (*labelling*), serta pembuangan *stopwords* (*stopword removal*).

Langkah awal dari pra-pemrosesan adalah tokenisasi untuk memotong kalimat menjadi kata. Pemotongan kata dilakukan berdasarkan spasi. Langkah selanjutnya adalah *case folding* kata yang sudah ditokenisasi. Kemudian, kata diperiksa (*word checking*) untuk mengurangi kata-kata yang memiliki kesalahan pengetikan atau singkatan untuk mengurangi daftar kata. Setelah itu, *token* dengan karakter berulang akan dihapus menjadi satu karakter. Langkah selanjutnya adalah pembuangan *stopwords*, seperti “untuk”, “yang”, “wah”, dan lain sebagainya. Selain itu, karakter lain selain a-z dan label dibuang dan dianggap tidak memiliki sentimen, seperti nomor telepon, simbol-simbol yang berdiri sendiri, *emoticon*, dan lain sebagainya.

Pemberian label pada kata-kata yang bersifat menolak seperti “tidak”, “gak”, “ogah”, dan lain-lain diberi label “//NOT//” dan kata-kata yang bersifat penguat seperti “banget”, “sangat”, “bingits”, dan lain-lain diberi label “//EXCESS//”. Tanda seru akan diberi label “//!//” dan tanda tanya akan diberi label “//?//”.

Langkah berikutnya adalah ekstraksi fitur menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Setelah melakukan klasifikasi dengan masing-masing algoritma, jenis sentimen dari suatu *query* dapat diketahui. Tahap berikutnya adalah mengevaluasi akurasi dari masing-masing algoritma klasifikasi. Akurasi dari pengklasifikasi dilihat dengan menggunakan *confusion matrix*.

Hasil akhir dari sistem analisis sentimen adalah jenis sentimen yang dihasilkan. Evaluasi yang akan digunakan dalam penelitian ini adalah *confusion matrix* untuk mengevaluasi hasil seleksi data.

## HASIL DAN PEMBAHASAN

### Analisa Sistem

Salah satu cara menganalisis jenis-jenis sentimen adalah dengan menggunakan sistem analisis sentimen. Sentimen atau opini masyarakat dapat diperoleh dari kegiatan berkomunikasi, baik secara tertulis, maupun lisan. Salah satu kegiatan berkomunikasi secara tertulis adalah *microblogging*. Analisis sentimen bertujuan untuk melakukan ekstraksi atribut dan komponen dari opini tersebut serta menentukan apakah opini tersebut bersifat positif atau negatif.

Sistem analisis sentimen dibutuhkan untuk mendapatkan sentimen dari suatu topik, misalnya nama produk, tokoh, negara, agama, dan lain-lain. Dengan sistem analisis sentimen, seseorang dapat

mengetahui sentimen yang diberikan oleh para pengguna *microblog* terhadap suatu topik pada saat itu.

### Implementasi Sistem WebAdminLogin

WebAdminLogin merupakan tampilan awal dari situs IndoSen. Jenis pengguna situs Indosen ada 2, yaitu *admin* dan *guest*. *Guest* tidak perlu memasukkan *password*, sedangkan *admin* perlu memasukan *username* dan *password*. Klik tombol “Masuk” untuk masuk ke dalam halaman khusus pengguna.



Gambar 1. Halaman Utama

### HALAMAN GUEST WebLoginGuest

Pencarian kata kunci yang ingin diketahui sentimen pada *microblog* ada pada halaman khusus *guest*. Masukkan kata kunci yang ingin diketahui sentimennya, misalnya “jokowi”, “indonesia”, “jakarta”, dan lain-lain. Klik tombol “Cari”. Tombol “Tentang” berisi keterangan situs IndoSen. Tombol “Kembali” untuk kembali ke halaman utama.



Gambar 2. Halaman Utama Guest

### WebResultGuest

Hasil akan ditampilkan dalam bentuk *pie chart* dan tabel. *Pie chart* menampilkan persentase dari masing-masing kelas. Tabel menampilkan kalimat hasil pencarian kata kunci dengan sentimennya. Untuk melihat hasil sentimen dengan algoritma lainnya, dapat menggunakan 5 tombol metode yang ada. Tombol “Kembali” untuk kembali ke halaman utama *guest*.



Gambar 3. Hasil Klasifikasi dalam Bentuk *Pie Chart*

*Pie chart* berisi persentase kelas pada suatu metode dengan warna kelas.

NO	DATA	KELAS
1	Jahak yg tau	Sangat Positif
2	pak jenderal tau ya untuk permasalahan DKK?	Netral
3	Kan mantan gubernur	Negatif
4	pak ky bisa di situ atau bisa diru malah banyak praktik kor dan dana dia yg sudah 3 kali cari di dia ya ky agi dia yg milihnya	Negatif
5	pk presiden smaha mppg	Negatif
6	dan utk Pak Jha, dan bisa utk itu sebagai bukannya, yg ke 70, amoga itu bisa sebagai juga, Amah YGA	Positif
7	hanya pendatang dan yg bilang negara baik saja	Negatif
8	STPN Pertanyaa Agas Lalu Penda di NTS	Negatif
9	0: "Berada dalam yg malingnya alam, bukan hankam"	Positif
10	Begitupun dgn masalah pemangin	Negatif
11	I Wed a video Gubernur na kudu akses di depan umum, perbaiki apa yang terjadi	Positif
12	Web Harapan Prastaban Bogor "Berbuka 2017, Presiden Jokowi Perlu Konsentrasi Tinggi Untuk Peroleh Target"	Positif
13	KEMANA JORONG PERUSAHAAN BIKIN? GARA KEMAHARAHAN TIKUAP? (Jawaban)	Negatif

Gambar 4. Hasil Klasifikasi dalam Bentuk Tabel

Tabel hasil klasifikasi berisi jumlah data yang terambil, kalimat dari data tersebut, dan kelas dari hasil pengklasifikasi.

### HALAMAN ADMIN WebHome

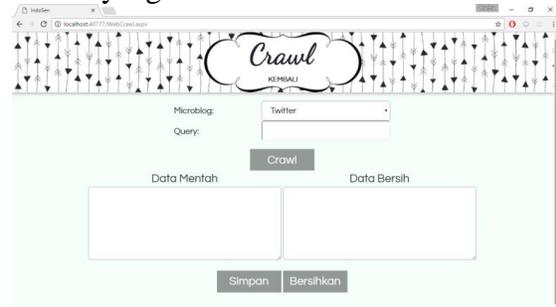
WebHome merupakan halaman utama *admin*. Pada halaman ini, *admin* dapat melakukan penambahan data pada tombol “Crawl”, pelatihan data pada tombol “Latih”, dan pengujian data pada tombol “Uji”. Fungsi tombol “Tentang” dan tombol “Kembali” sama dengan fungsi tombol yang ada pada halaman utama *guest*.



Gambar 5. Halaman Utama Admin

### WebCrawl

WebCrawl digunakan untuk melakukan penambahan data untuk dilatih. Pilih *microblog* yang diinginkan pada “Microblog”. Kemudian, masukkan *query* untuk Twitter, *page name* untuk Facebook, atau *username* untuk YouTube. Klik tombol “Crawl”. Data mentah dan data bersih akan ditampilkan pada masing-masing *textbox*. Klik tombol “Simpan” untuk menyimpan data hasil *crawl*. Klik “Bersihkan” untuk menghapus seluruh kolom yang ada.



Gambar 6. Halaman Penambahan Data Siap Latih

### WebTrain

WebEdit digunakan untuk melatih data siap latih dan melakukan perubahan terhadap data latih yang telah tersimpan. Tombol “Tambah” untuk kel halaman tambah data latih. Klik *microblog* yang telah dilakukan proses *crawl* sebelumnya. Kemudian, kalimat dari data siap latih akan muncul pada *textbox* yang berwarna abu-abu. Klik tombol kelas yang sesuai dengan kalimat yang ditampilkan. Apabila tidak sesuai, klik tombol “Bukan Sentimen” untuk membuang data tersebut. Apabila terdapat kesalahan dalam melatih, klik tombol “Edit” untuk masuk ke halaman WebEdit.



Gambar 7. Halaman Pelatihan Data

### WebEdit

WebEdit digunakan untuk memperbaiki data latih. Kolom kelas menampilkan kelas yang telah dilatih sebelumnya. Klik kelas dari data tersebut untuk meng-*edit* kelas dari data yang dipilih. Pilih kelas baru dan kelas dari data tersebut akan otomatis diperbarui. Klik tombol “Hapus” pada baris data yang ingin dihapus untuk membuang data tersebut dari data latih.



Gambar 8. Halaman untuk Memperbaiki Data Latih

### WebTest

WebTest merupakan pengujian otomatis sistem dengan memilih jumlah data yang ingin dijadikan data uji dan jumlah K tetangga. Data latih yang dijadikan data uji akan dibuang sementara pada penyimpanan data latih. Data tersebut akan dikembalikan apabila klasifikasi data tersebut telah selesai. Tombol “Kalimat” untuk melakukan pengujian terhadap kalimat yang diinginkan. Tombol “Kembali” untuk kembali ke menu halaman utama *admin*.

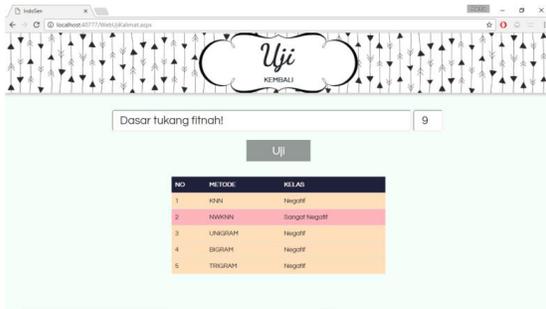


Gambar 9. Halaman Pengujian Data

### Web Uji Kalimat

Halaman Web Uji Kalimat digunakan untuk menguji kalimat yang diinginkan. Masukkan kalimat yang ingin diuji pada kolom “Kalimat” sebelah kiri dan masukkan jumlah K tetangga metode KNN dan NW-KNN. Klik tombol “Uji” dan tabel akan hasil akan menampilkan kelas hasil klasifikasi dari masing-masing metode. Sebagai contoh, kalimat yang dimasukkan adalah “Dasar Tukang

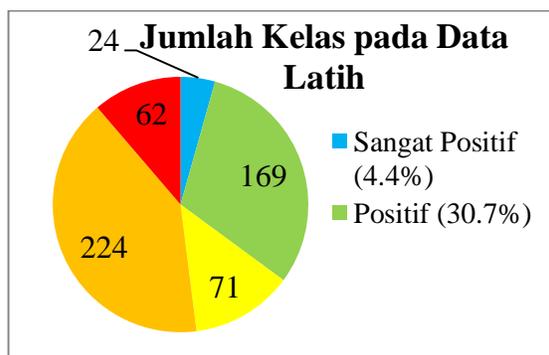
Fitnah!” dengan  $K = 9$ . KNN mengklasifikasi kalimat tersebut dalam kelas negatif. Sedangkan NW-KNN mengklasifikasi kalimat tersebut dalam kelas sangat negatif.



Gambar 1. Halaman Pengujian Kalimat

### Pengujian Sistem

Data latih yang digunakan sejumlah 550 data dengan komposisi kelas yang tidak seimbang. Data latih yang digunakan dapat dilihat pada Gambar 12.



Gambar 2. Jumlah Kelas pada Data Latih

Jumlah kelas terbanyak ada pada kelas negatif dengan 224 data (40.7%), diikuti dengan kelas positif sejumlah 169 data (30.7%), kelas netral sejumlah 71 data (12.9%), sangat negatif sejumlah 62 data (11.3%), dan yang terakhir adalah kelas sangat positif sejumlah 24 data (4.4%).

Seluruh percobaan menggunakan 50 data uji. Data uji tersebut dibagi menjadi 10 data untuk masing-masing jenis kelas yang ada. Evaluasi dari masing-masing

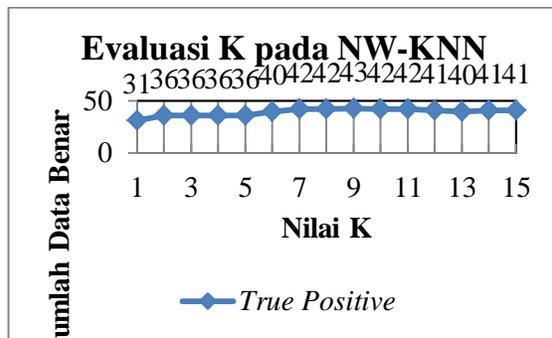
metode menggunakan *confusion matrix* dengan mengambil nilai *true positive* dan akurasi.

Percobaan yang dilakukan adalah percobaan metode KNN dan NW-KNN. Akurasi dari pengklasifikasi NW-KNN didapatkan sebesar 86% dengan menggunakan  $K = 9$ . Hasil akurasi NW-KNN dapat dilihat pada Tabel 3.

Tabel 3. Hasil NW-KNN dengan  $K = 9$

NO	KELAS	BENAR	AKURASI
1	Sangat Positif	9	90%
2	Positif	9	90%
3	Netral	10	100%
4	Negatif	7	70%
5	Sangat Negatif	8	80%
<b>TOTAL</b>		43	86%

NW-KNN dapat mengklasifikasi 9 data dari 10 data dengan kelas sangat positif, 9 data dari 10 data dengan kelas positif, 10 data dari 10 data dengan kelas netral, 7 data dari 10 data dengan kelas negatif, dan 8 data dari 10 data dengan kelas sangat negatif. Total data yang dapat diklasifikasi dengan benar adalah 43 data dari 50 data. Dengan demikian, akurasi dari metode NW-KNN pada pengujian yang dilakukan sebesar 86%. Contoh kesalahan misalnya pada data uji “Ah ini mah propagandanya timses tetangga”. Data seharusnya masuk dalam kelas negatif. Akan tetapi, NW-KNN mengklasifikasi data ke dalam kelas sangat negatif. Gambar 12 merupakan tabel evaluasi K pada NW-KNN.



Gambar 3. Evaluasi K pada NW-KNN

Nilai K dengan akurasi tertinggi pada NW-KNN adalah pada saat K = 9. Pada saat K = 9, NW-KNN dapat mengklasifikasi 43 data dari 50 data secara benar. Dan nilai K dengan akurasi terendah adalah pada saat K = 1. Pada saat K = 1, NW-KNN hanya dapat mengklasifikasi 31 data dari 50 data secara

benar. Akurasi dari NW-KNN naik pada saat K = 6. Kesalahan klasifikasi data uji pada NW-KNN dapat dilihat pada Tabel 2.

Kesalahan klasifikasi pada NW-KNN dapat dilihat pada Tabel 4. Kesalahan terbanyak ada pada saat kelas latih seharusnya masuk ke kelas negatif. Hal ini dapat dipengaruhi oleh kata dalam data latih. Misalnya, sistem mengklasifikasikan data pada Tabel 2 Nomor 7 dengan kelas 5 (sangat positif) ke dalam kelas 4 (positif). Pada data latih, kata “cantik” terdapat pada data latih “Dan kota makin terlihat cantik.” dan “Buat strategi cantik bu,spy dapat.” Kedua data latih tersebut diberi kelas 4 (positif) pada sistem.

Tabel 4. Kesalahan Klasifikasi pada NW-KNN

NO	KALIMAT UJI	KELAS	
		LATIH	UJI
1	Disaat Ahok berkuasa, Ahok melihatkan kesombongannya,tapi saat tidak berkuasa kelihatan kepegecutannya.	1	2
2	se-Pandai2nya Manusia Menyimpan Bangkai, Pasti Keciuman.	1	5
3	Lho? Kok jadi kmu yg ngatur???	2	3
4	sukurin senjata makan tuhan...	2	3
5	Ahh ini mah propagandanya timses tetangga	2	1
6	seneng deh liatnya...	4	1
7	cantik banget anak satu ini!	5	4

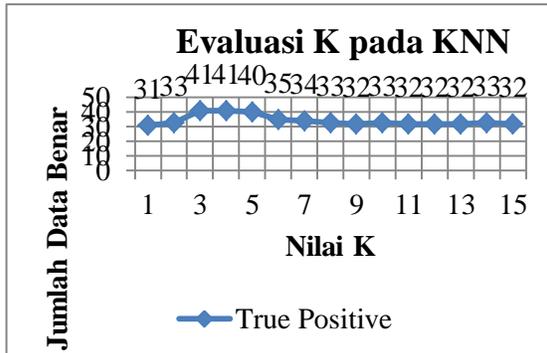
Percobaan pada pengklasifikasi KNN memiliki akurasi sebesar 82% dengan menggunakan K = 3. Hasil akurasi KNN dapat dilihat pada Tabel 5.

Tabel 5. Hasil KNN dengan K = 3

NO	KELAS	BENAR	AKURASI
1	Sangat Positif	5	50%
2	Positif	10	100%
3	Netral	10	100%
4	Negatif	9	90%
5	Sangat Negatif	7	70%
TOTAL		41	82%

KNN dapat mengklasifikasi 5 data dari 10 data dengan kelas sangat positif, 10 data dari 10 data dengan kelas positif, 10 data dari 10 data dengan kelas netral, 9 data dari 10 data dengan kelas negatif, dan 7 data dari 10 data dengan kelas sangat negatif. Total data yang dapat diklasifikasi dengan benar adalah 41 data dari 50 data. Akurasi KNN sebesar 82%. Contoh kesalahan misalnya pada data uji “Lho? Kok jadi kmu yg ngatur???”. Data tersebut seharusnya masuk ke kelas negatif. Akan tetapi, KNN mengklasifikasi data ke kelas

netral. Gambar 13 merupakan tabel evaluasi K pada KNN.



Gambar 4. Evaluasi K pada KNN

Nilai K dengan akurasi tertinggi pada KNN adalah pada saat K = 3. Pada saat K = 3, KNN dapat mengklasifikasi 41 data dari 50 data secara benar. Dan nilai K dengan akurasi terendah adalah pada saat K = 1. Pada saat K = 1, KNN hanya dapat mengklasifikasi 31 data dari 50 data secara benar. Akurasi dari KNN turun pada saat K = 6. Kesalahan klasifikasi data uji pada KNN dapat dilihat pada Tabel 6.

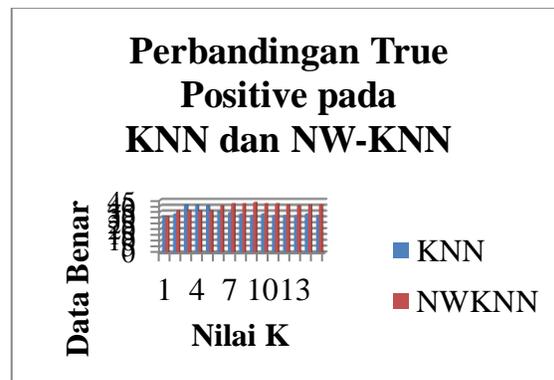
Tabel 6. Kesalahan Klasifikasi pada KNN

NO	KALIMAT UJI	KELAS	
		LATIH	UJI
1	Disaat Ahok berkuasa, Ahok melihat kesombongannya,tapi saat tidak berkuasa kelihatan kepengecutannya.	1	2
2	se-Pandai2nya Manusia Menyimpan Bangkai, Pasti Keciuman.	1	5
3	kali penuh sampah.	1	4
4	Lho? Kok jadi kmu yg ngatur???	2	3
5	Pdhl jakarta udah lbh bersih dari dlu	5	2
6	Dan kota makin terlihat cantik.	5	4
7	cantik banget anak satu ini!	5	4
8	rindu banget nasi impit rendang kuah kacang!!	5	2
9	Qlue hebat banget, jadi iri sama Jakarta.	5	1

Kesalahan terbanyak saat kelas seharusnya masuk ke kelas sangat positif. Hal ini dapat dipengaruhi oleh kata dalam data latih. Misalnya, sistem mengklasifikasikan data pada Tabel 3 Nomor 7 dengan kelas 5 (sangat positif) ke kelas 4 (positif). Pada data latih, kata “cantik” terdapat pada data “Dan kota makin terlihat cantik.” dan “Buat strategi cantik bu,spy dapat.” Kedua data diberi kelas 4 (positif) pada sistem.

Dengan pembobotan kelas, NW-KNN memiliki akurasi yang lebih tinggi dibandingkan dengan KNN yang tidak melakukan pembobotan kelas. Gambar 15

menunjukkan perbandingan nilai *true positive* antara KNN dan NW-KNN.



Gambar 5. Perbandingan *True Positive* pada KNN dan NW-KNN

Sebagian besar nilai *true positive* dari NW-KNN lebih tinggi dibandingkan KNN. Nilai *true positive* KNN yang lebih besar dibandingkan NW-KNN hanya pada saat  $K = 3$ ,  $K = 4$ , dan  $K = 5$ .

Dengan demikian, pada percobaan ini, pembobotan kelas yang dilakukan oleh metode NW-KNN dapat meningkatkan akurasi pada data dengan komposisi kelas yang tidak seimbang seperti yang dimiliki oleh data latih percobaan. Akurasi tertinggi KNN didapatkan sebesar 82% dengan  $K = 3$ , sedangkan akurasi tertinggi NW-KNN didapatkan sebesar 86% dengan  $K = 9$  menggunakan data latih dan data uji yang sama.

## PENUTUP

### Kesimpulan

1. Pada penelitian ini, metode NW-KNN memiliki akurasi sebesar 86% dengan menggunakan data latih sejumlah 550 dan menggunakan  $K = 9$ .
2. Hasil dari pengujian yang dilakukan menunjukkan bahwa metode NW-KNN dapat mengatasi permasalahan data yang tidak seimbang pada data latih sejumlah 550.
3. Nilai  $K$  pada NW-KNN dapat mengklasifikasi data paling baik pada saat  $K = 9$  pada data latih yang berjumlah 550 dan KNN pada saat  $K = 3$  pada data latih yang sama pada pengujian yang dilakukan.

## DAFTAR PUSTAKA

Ehlers, U. D. (2013). *Open learning cultures. A Guide to Quality, Evaluation, and Assessment for Future Learning*. Heidelberg, Berlin.

Grossman, D. A., & Frieder, O. (2012). *Information retrieval: Algorithms and heuristics* (Vol. 15). Springer Science & Business Media.

Harlili, & Wibisono, Yudi (2013). *Sistem Analisis Opini Microblogging Berbahasa Indonesia*. Departemen Teknik Informatika Institut Teknologi Bandung.

Indriati, I., & Ridok, A. (2016). *Sentiment Analysis For Review Mobile Applications Using Neighbor Method Weighted K-Nearest Neighbor (Nwknn)*. Journal of Environmental Engineering and Sustainable Technology, 3(1), 23-32.

Jurafsky, D., & James, H. (2000). *Speech and language processing an introduction to natural language processing, computational linguistics, and speech*.

Keller, K. L., & Lehmann, D. R. (2006). *Brands and branding: Research findings and future priorities*. *Marketing science*, 25(6), 740-759.

Retnawiyati, Eka, et al. (2015). *Analisis Sentimen pada Data Twitter dengan Menggunakan Text Mining terhadap Suatu Produk*. Departemen Teknik Informatika Universitas Bina Darma.

Ridok, A., & Latifah, R. (2015). *Klasifikasi Teks Bahasa Indonesia Pada Corpus Tak Seimbang Menggunakan Nwknn*. *Proceedings Konferensi Nasional Sistem dan Informatika (KNS&I)*.

Suprpto, F. (2015). *Pengembangan Aplikasi Sentiment Analysis Menggunakan Metode Naïve Bayes (Studi Kasus Sentiment Analysis Dari Media Twitter)*. SESINDO 2015.