p-ISSN: 2339-1928 & e-ISSN: 2579-633X

Penerapan Algoritma K-Nearest Neighbors Pada Analisis Sentimen Review Agen Travel

Siti Ernawati[1], Risa Wati[2]
[1]STMIK Nusa Mandiri Jakarta
e-mail: rna2103@gmail.com
[2]AMIK BSI Tasikmalaya
e-mail: risawati06@gmail.com

ABSTRAKSI

Penggunaan internet mampu memberikan pengaruh terhadap orang-orang melalui internet; kita bisa mendapat informasi, selain itu, kita juga bisa memberi pendapat positif dan negatif untuk review tertentu. Dengan menyediakan banyak data atau informasi di internet, kami menggunakannya untuk diproses jadi, itu akan memiliki pengetahuan baru. Berdasarkan hal itu, penulis membuat penelitian, seperti klasifikasi opini dengan menganalisis sentimen melalui pendekatan penambangan teks, dalam penelitian ini; dibutuhkan suatu metode yang mampu mengklasifikasikan pendapat secara akurat. Ruang lingkup penelitian ini adalah peninjauan agen perjalanan pengolahan data menggunakan algoritma K-Nearest Neighbor (K-NN) yang menggunakan 100 review positif dan 100 review negatif dengan enam kata yang berhubungan dengan sentimen yaitu: Fast, Good, Great, Buruk, Cencel, dan Tunggu. Ini memiliki bukti bahwa dengan menggunakan algoritma K-Nearest Neighbor (K-NN), ia mencapai hasil akurasi terbaik dan berdasarkan perhitungan yang dinyatakan dalam aplikasi. Titik akurasi peninjauan agen perjalanan menggunakan K-Nearest Neighbor (K-NN) algoritma telah mencapai 87,00% dan titik AUC adalah 0,916, titik AUC milik kelompok Klasifikasi Excellent sehingga dinyatakan bahwa K-Nearest Neighbor (K -NN) memiliki hasil yang akurat dalam menganalisis sentimen ulasan agen perjalanan.
Kata kunci: **Analisis sentimen, K-Nearest Neighbor dan Travel Agent Review**

ABSTRACT

The internet use is able to give the affect toward people through internet; we can get information, besides that, we can also give positive and negative opinion for a certain review. By providing a lot of data or information on the internet, we use it to be processed so, it will have new knowledge. Based on that matter, the writer makes a research, such as the classification of opinion by analyzing sentiment through the approach of text mining, within this research; it needs a method that is able to classify opinion accurately. The scope of this research is the travel agent review of data processing uses K-Nearest Neighbor (K-NN) algorithm which uses 100 positive reviews and 100 negative reviews by six words which is related to sentiment those are: Fast, Good, Great, Bad, Cencel and Wait. It has evidence that by using K-Nearest Neighbor (K-NN) algorithm, it reaches the best accuracy result and based on the stated calculation in an application. The point of travel agent review accuracy uses K-Nearest Neighbor (K-NN) algorithm has reached 87.00% and the point of AUC is 0.916, point of AUC belongs to Excellent Classification group so that it is stated that K-Nearest Neighbor (K-NN) has the accurate result in analyzing travel agent review sentiment.
Keyword: ***Sentiment analysis, K-Nearest Neighbor and Travel Agent Review***

1. PENDAHULUAN

Meningkatnya penggunaan internet akan memberikan dampak terhadap kehidupan masyarakat di dunia, melalui internet masyarakat bisa mendapatkan banyak sekali informasi mengenai hal apapun yang sedang

terjadi. Selain mendapatkan informasi, dalam dunia internet pun kita dapat menyampaikan opini atau saran terhadap informasi terkini. Opini yang diberikan bisa dalam bentuk opini positif ataupun opini negatif. Berdasarkan banyaknya ketersediaan data yang ada di internet kita bisa memanfaatkan data tersebut untuk diolah sehingga menghasilkan pengetahuan yang baru.

Dalam hal pengolahan data, peneliti akan meneliti bagaimana mengklasifikasikan opini yang diberikan masyarakat terhadap agen travel, baik opini terhadap biaya dan pelayanan travel. Proses pengklasifikasian opini ini berupa analisis sentimen melalui pendekatan text mining sehingga diperlukan metode yang mampu mengklasifikasikan opini secara akurat. Dalam pengklasifikasian ini menggunakan metode K-Nearest Neighbor (K-NN) dan pembobotan untuk setiap kata menggunakan TF-IDF. Ada beberapa kelebihan dari metode K-NN yaitu algoritma klasifikasi K-Nearest Neighbor yang terbukti mencapai hasil akurasi yang baik dan sesuai dengan perhitungan yang diterapkan dalam sebuah aplikasi (Sani, Zeniarza, & Luthfiarta, 2016). Kinerja K-NN sebagai algoritma klasifikasi cukup bagus ditunjukkan oleh beberapa penelitian yang menggunakannya (Sani, Zeniarza, & Luthfiarta, 2016). Algoritma K-Nearest Neighbor sangat umum digunakan untuk pengkategorisasian teks (Samuel, Delima, & Rachmat, 2014).

Hal tersebut diketahui karena algoritmanya yang mudah dan efisien untuk klasifikasi teks. Bukan hanya mudah dan efisien, sifat dari algoritma K-Nearest Neighbor sendiri bersifat *self-learning*, dimana algoritma ini dapat mempelajari struktur data yang ada dan menkategorikan dirinya sendiri (Samuel, Delima, & Rachmat, 2014).

Selain itu terdapat beberapa penelitian sebelumnya dalam hal klasifikasi sentimen yaitu sentimen analisis pada teks Bahasa Indonesia menggunakan Support Vector Machine (SVM) dan K-Nearest Neighbor (Lidya, Sitompul, & Efendi, 2015), klasifikasi artikel wikipedia Bahasa Indonesia menggunakan metode K-Nearest Neighbor (Hardiyanto, & Rahutomo, 2016) dan penentuan Tugas Akhir berdasarkan data abstrak menggunakan algoritma K-Nearest Neighbor (Sani, Zeniarza, & Luthfiarta, 2016).

2. TINJAUAN PUSTAKA

2.1. Text Mining

Text Mining adalah proses pengetahuan intensif dimana pengguna berinteraksi koleksi dokumen dari waktu kewaktu dengan menggunakan rangkaian alat analisis (Ronen, & James, 2007). Text mining juga memiliki definisi menambang data berupa teks dimana sumber data biasanya didapatkan dari dokumen, dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dukumen.

2.2. Analisis Sentimen (*Sentiment Analysis*)

Analisis Sentimen adalah tugas mengidentifikasi opini positif dan negatif, emosi dan evaluasi (Wilson, Wiebe, & Hoffmann, 2005). Analisis Sentimen atau opini mining adalah studi komputasi tentang pendapat, sentimen dan emosi yang dinyatakan dalam teks (Bing, 2010).

Langkah-langkah umum pada analisis sentimen klasifikasi teks adalah

1. Definisikan domain dataset
   Mengumpulkan dataset seperti *review* agen trevel, *review* restoran, *review* produk dan lain-lain
2. *Pre-processing*
   Pada tahapan ini biasanya dilakukan *Tokenization*, *Stopwords Removal*, dan *Stemming*.
3. *Transformation*
   Pembobotan dari data tekstual, proses yang sering digunakan adalah TF-IDF.
4. *Feature Selection*
   Membuat pengklasifikasi lebih efisien dengan mengurangi jumlah data yang dianalisa.
5. *Classification*
   Pengklasifikasi teks biasanya menggunakan metode Naive Bayes, K-Nearest Neighbor, SVM dan lain-lain.
6. *Interpretation/Evaluation*
   Biasanya evaluasi untuk menghitung nilai akurasi dan nilai AUC.

2.3. K-Nearest Neighbor

Algoritma K-Nearest Neighbor merupakan sebuah algoritma yang sering

digunakan untuk klasifikasi teks dan data (Samuel, Delima, & Rachmat, 2014).

Tujuan dari algoritma ini adalah mengklasifikasikan obyek berdasarkan atribut dan *training sample*. *Clasifier* tidak menggunakan apapun untuk dicocokkan dan hanya berdasarkan pada memori. Diberikan titik *query*, akan ditemukan sejumlah k obyek atau (titik *training*) yang paling dekat dengan titik *query*. Klasifikasi menggunakan *voting* terbanyak diantara klasifikasi dari k obyek. Algoritma K-Nearest Neighbor (K-NN) menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari *query instance* yang baru (Lidya, Opim, & Syahril, 2015).

2.4. Perancangan Penelitian

Pada penelitian ini menggunakan penelitian eksperimen dengan tahapan sebagai berikut:

1. Pengumpulan Data Set
   Penelitian ini diawali dengan melakukan pengumpulan data. Data yang diperoleh dari laman yang berupa kumpulan opini masyarakat yang sudah banyak tersedia. Kemudian dari kumpulan opini tersebut diintegrasikan kedalam dataset.
2. Pengolahan Awal Data
   Setelah pengumpulan data langkah selanjutnya adalah pengolahan data, peneliti mengambil sampel sebanyak 100 *review* positif dan 100 *review* negatif sebagai data training. Pada pengolahan awal data melalui 3 proses yaitu:

a. *Tokenization*
   Tokenization adalah mengumpulkan semua kata dan menghilangkan tanda baca maupun simbol yang bukan huruf, seperti " , . /. ;) dan lain-lain.

b. *Stopwords Removal*
   Stopwords dapat diartikan sebagai menghilangkan kata-kata umum yang tidak memiliki makna atau informasi yang dibutuhkan, seperti *the, of, with, for* dan lai-lain.

c. *Stemming*
   Stemming merupakan salah satu proses dari mengubah token yang berimbuhan menjadi kata dasar, dengan menghilangkan semua imbuhan yang ada pada token tersebut. Pentingnya *stemming* dalam proses pembuatan sistem adalah untuk menghilangkan imbuhan pada awalan dan akhiran.

d. Metode yang diusulkan
   Pada penelitian ini, metode yang diusulkan adalah metode K-Nearest Neighbor yang terbukti mencapai hasil akurasi yang baik dan sesuai dengan perhitungan yang diterapkan dalam sebuah aplikasi.

e. Eksperimen dan Pengujian Metode
   Pada penelitian ini proses eksperimen menggunakan RapidMiner 5.3. Data training yang digunakan adalah *review* agen travel yang dikumpulkan dari situs https://www.trustpilot.com/categories/travel_holidays dikelompokkan menjadi 2 yaitu *review* positif dan *review* negatif.

3. Evaluasi dan Validasi Hasil Evaluasi
   Pada penelitian ini validasi dilakukan dengan menggunakan 10 *fold cross validation*. Akurasi diukur dengan *confusion Matrix*. Kurva ROC digunakan untuk mengukur nilai AUC.
   Panduan untuk mengklasifikasikan keakuratan diagnosa menggunakan AUC, disajikan dibawah ini (Gorunescu, 2011):

1. 0.90-1.00 = *Excellent Classification*;
2. 0.80-0.90 = *Good Classification*;
3. 0.70-0.80 = *Fair Classification*;
4. 0.60-0.70 = *Poor Classification*;
5. 0.50-0.60 = *Failure*.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Pada penelitian ini menggunakan data *review* agen travel yang dikumpulkan dari situs https://www.trustpilot.com/categories/travel_holidays, data terdiri dari 100 *review* positif dan 100 *review* negatif.

Berikut contoh *review* positif dan *review* negatif pada *review* agen travel :

1 *Review* Positif
Bought 3 flights, Sky-Tours ticketed only 2. Fortunately I checked and noticed. Due to my sending of proof of their confirmation of my booking and the difference with the ticket received, then they finally issued the missing flight, without ackwoledging their mistake. I wrote to them seeking clarification but got no

response. Not to be trusted.

2 *Review* Negatif

We were trying to book a flight which at the time ssid it was $888. When we selected it and tried to continue, it said that those tickets were sold out, and recommended other tickets for the same price, only the flight would be a bit longer. When we selected those, it told us the same message.My friend was also tryig to book at the same time as me and selected his flight, and it said that there were technical difficulties, and asked him if he wanted to wait, and someone would contact him (I believe it said within an hour) so that he could book the flight. He selected yhe wait option however nobody has contacted him.Eventualy we gave up and paid $100 more just to book the damn ticket. Hope all goes well with these tickets, as Ive been reading sketchy reviews now.Wouldnt recommend to a friend.

3.2 Pengolahan Data Awal

Pada tahap pengolahan data, data yang sudah terkumpul, terlebih dahulu diolah dengan melakukan penyeleksian data yang meliputi pembersihan data, mentransformasikan data kedalam bentuk yang dibutuhkan.

3.3 Metode yang Diusulkan

Setelah melakukan pengumpulan dan pengolahan data tahap selanjutnya adalah menentukan metode, dimana metode ini merupakan gambaran dari rangkaian kegiatan dan membagi data kedalam data *training* dan data *testing*.

3.4 Eksperimen dan Pengujian Metode

Tahapan dasar yang dilakuakan dalam proses kalsifikasi yaitu:

1. *Tokenizing*

Tabel 1. *Tokenizing*

Teks Sebelum Proses Tokenization	Teks Setelah Proses Tokenization
Bought 3 flights, Sky-Tours ticketed only 2. Fortunately I checked and noticed. Due to my sending of proof of their confirmation of my booking and the difference with the ticket received, then they finally issued the missing flight, without ackwoledging their mistake. I wrote to them seeking clarification but got no response. Not to be trusted.	*Bought flights Sky Tours ticketed only Fortunately I checked and noticed Due to my sending of proof of their confirmation of my booking and the difference with the ticket received, then they finally issued the missing flight, without ackwoledging their mistake I wrote to them seeking clarification but got no response Not to be trusted*

Sumber: Peneliti (2018)

2. *Stopwords Removal*

Tabel 2. *Stopwords Removal*

Teks Sebelum Proses *Stopwords Removal*	Teks Setelah Proses *Stopwords Removal*
Bought 3 flights, Sky-Tours ticketed only 2. Fortunately I checked and noticed. Due to my sending of proof of their confirmation of my booking and the difference with the ticket received, then they finally issued the missing flight, without ackwoledging their mistake. I wrote to them seeking clarification but got no response. Not to be trusted.	*Bought flights Sky Tours ticketed Fortunately I checked noticed Due sending proof confirmation booking difference ticket received finally issued missing flight ackwoledging mistake I wrote seeking clarification got response trusted*

Sumber: Peneliti (2018)

3. *Stemming*

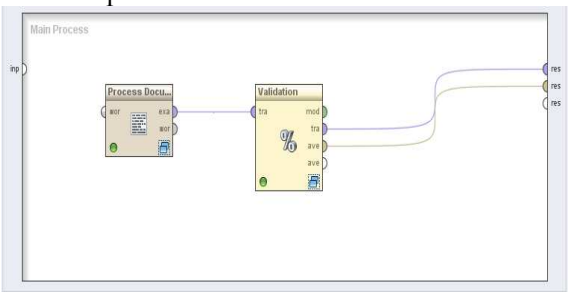
Tabel 3. *Stemming*

Teks Sebelum Proses *Stemming*	Teks Setelah Proses *Stemming*
Bought 3 flights, Sky-Tours ticketed only 2. Fortunately I checked and noticed. Due to my sending of proof of their confirmation of my booking and the difference with the ticket received, then they finally issued the missing flight, without ackwoledging their mistake. I wrote to them	*bought flight sky tour ticket fortun i check notic due send proof confirm book differ ticket receiv final issu miss flight ackwoledg mistak i wrote seek clarif got respons trust*

seeking clarification but got no response. Not to be trusted.	

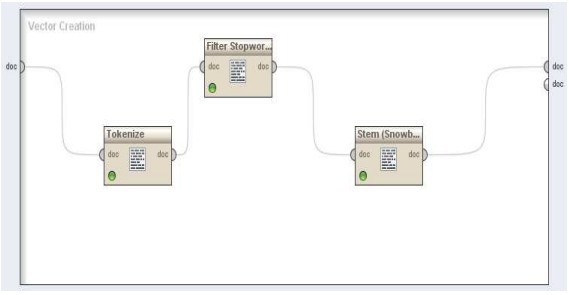
Sumber: Peneliti (2018)

Saat melakukan eksperimen menggunakan metode K-Nearest Neighbor peneliti menggunakan data sebanyak 200 *review*. Selain itu juga peneliti menggunakan *10-fold cross validation* untuk pengujian model, dimana setiap bagian akan dibentuk secara random. Prinsip *10-fold cross validation* adalah 1:9, 1 bagian menjadi data *testing* dan data lainnya menjadi data *training*, sehingga 10 bagian tersebut berkesempatan menjadi data *testing*.
Berikut adalah tampilan pemrosesan data dalam rapidminer:



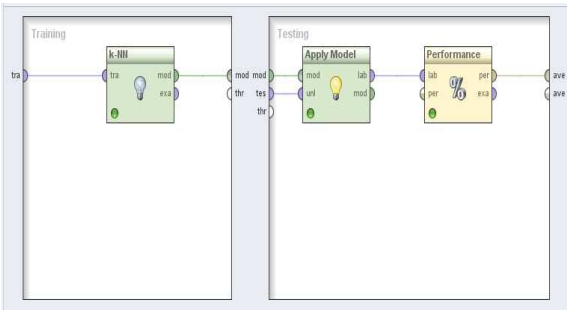
Sumber: Peneliti (2018)

Gambar 1. Tampilan saat memasukan nilai validasi



Sumber: Peneliti (2018)

Gambar 2. Tampilan saat proses preprosesing



Sumber: Peneliti (2018)

Gambar 3. Tampilan saat memasukan metode K-NN

Berikut adalah tabel eksperimen dengan merubah nilai dari K:

Tabel 4. Eksperimen dengan merubah nilai K

Nilai K	Accuracy	Precision	Recall	AUC
1	80.00	77.88	84.00	0.500
2	82.50	89.82	74.00	0.823
3	80.50	78.48	85.00	0.874
4	82.00	84.37	79.00	0.894
5	81.50	79.77	85.00	0.907
6	83.50	88.39	78.00	0.908
7	84.00	85.54	83.00	0.910
8	**87.00**	**92.94**	**81.00**	**0.916**
9	85.00	86.05	84.00	0.936
10	84.50	87.72	81.00	0.909
11	86.50	87.16	86.00	0.912
12	85.50	89.51	81.00	0.914
13	86.00	87.71	84.00	0.912
14	85.00	89.98	79.00	0.911
15	83.50	85.08	81.00	0.912
16	83.00	88.74	76.00	0.914
17	82.50	85.74	78.00	0.906
18	83.00	89.88	75.00	0.914
19	83.00	87.70	77.00	0.915
20	82.50	88.45	75.00	0.914

Sumber: Peneliti (2018)

Berdasarkan perubahan nilai K yang sudah dilakukan, maka hasil *accuracy* tertinggi berada pada posisi nilai K=8 dengan nilai *accuracy* sebesar **87.00%** dan **nilai AUC sebesar 0.916.**

Tabel 5. Hasil akurasi menggunakan algoritma *K-Nearest Neighbor*

Accuracy: 87.00% +/- 3.32% (mikro: 87.00%)			
	true negatif	*true* positif	*class precision*
pred.negatif	93	19	83.04%
pred.positif	7	81	92.05%
class recall	93.00%	81.00%	

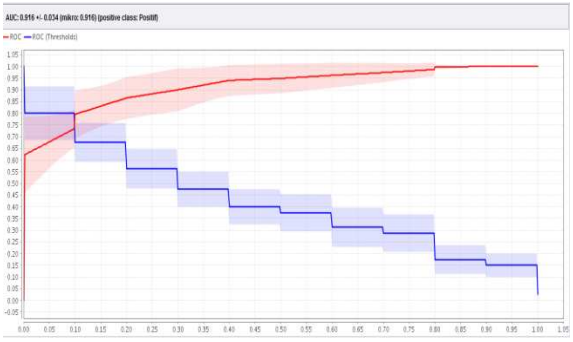
Sumber: Peneliti (2018)

Nilai *accuracy* dari *confusion matrix* tersebut adalah:

$$Accuracy = \frac{(TN + TP)}{(TN + FN + TP + FP)}$$

$$Accuracy = \frac{(93 + 81)}{(93 + 7 + 81 + 19)}$$

$$Accuracy = \frac{174}{200} = 0.87 = 87.00\%$$



Sumber: Peneliti (2018)

Gambar 4. Grafik *Area Under Curve (AUC)* menggunakan Metode *K-Nearest Neighbour*

Area Under Curve (AUC) yang dihasilkan adalah **0.916** dimana pengklasifikasian keakuratan dalam penelitian ini masuk kedalam kelompok *Excellent Classification.*

3.5 Evaluasi dan Validasi Hasil

Tahap evaluasi merupakan tahap akhir dari rangkaian kegiatan penelitian ini. Setelah melakukan tahap pengujian model maka akan menghasilkan nilai akurasi dan AUC. Kemudian dari hasil itu dievaluasi, dari hasil evaluasi itu dapat ditarik kesimpulan dari hasil penelitian ini.

4. KESIMPULAN

Penelitian ini melakukan pengklasifikasi teks *review* agen travel dengan pengklasifikasi K-Nearest Neighbor, menggunakan 100 *review* positif dan 100 *review* negatif serta enam kata yang berhubungan dengan sentimen yaitu *Fast, Good, Great, Bad, Cencel* dan *Wait*. Nilai akurasi yang dihasilkan mencapai 87.00% dengan nilai AUC sebesar 0.916 masuk kedalam kelompok *Excellent Classification.*

REFERENSI

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

Gorunescu, Florin (2011). Data Mining: Concepts, Models, and Techniques. Verlag Berlin Heidelberg: Springer.

Lidya, S. K., Sitompul, O. S., & Efendi, S. (2015). Sentiment Analysis Pada Teks Bahasa Indonesia Menggunakan Support Vector Machine (SVM) Dan K-Nearest Neighbor (K-NN). In *Seminar Nasional Teknologi Informasi dan Komunikasi*.

Liu, B. (2010). Sentiment Analysis and Subjectivity. *Handbook of natural language processing*, 2, 627-666

Lidya, K.,S, Opim, S., & Syahril, E. (2015). Sentiment Analysis Pada Text Bahasa Indonesia Menggunakan Support Vector Machine (SVM) dan K-Nearest Neighbor (K-NN). Seminar Nasional Teknologi Informasi dan Komunikasi 2015, Yogyakarta.

Nugroho, M. A., & Santoso, H. A. (2016). Klasifikasi Dokumen Komentar Pada Situs Youtube Menggunakan Algoritma K-Nearest Neighbor (K-NN). *Jurnal Sistem Informasi*.

Samuel, Y., Delima, R., & Rachmat, A. (2014). Implementasi Metode K-Nearest Neighbor dengan Decision Rule untuk Klasifikasi Subtopik Berita. Jurnal Informatika, 10(1), 1-15.

Sani, R. R., Zeniarza, J., & Luthfiarta, A. (2016). Pengembangan Aplikasi Penentuan Tema Tugas Akhir Berdasarkan Data Abstrak Menggunakan Algoritma K-Nearest Neighbor. Prosiding seminar Nasional Multi Disiplin Ilmu & Call For Papers, 103-111.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 347-354). Association for Computational Linguistics