Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 3, No. 5, Mei 2019, hlm. 4164-4169 http://j-ptiik.ub.ac.id

Klasifikasi Rating Berdasarkan Komentar Tempat Wisata Di Media Sosial Dengan Menggunakan Metode *Fuzzy K-Nearest Neighbor*

Nanda Ajeng Kartini[1], Fitra A Bachtiar[2], Indriati[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]najengkartini@gmail.com, [2]fitra.bachtiar@ub.ac.id, [3]indriati@ub.ac.id

Abstrak

Pada masa sekarang ini dengan kemudahan akses informasi, banyaknya website tempat wisata menggunakan rating fitur untuk membantu mempermudah informasi. Rating digunakan sebagai indikator untuk menunjang kualitas dan popularitas. Pengguna hanya memberi penilaian secara keseluruhan pada setiap komentar dan tidak memberikan penilaian sesuai dengan aspek yang dibicarakan, sehingga menyulitkan pembaca komentar menganalisis aspek yang unggul pada komentar tersebut. Dari masalah tersebut, pada penelitian ini akan dibuat suatu sistem klasifikasi rating pada tempat wisata dengan menggunakan metode *Fuzzy K-Nearest Neighbor* (FKNN). Metode FKNN merupakan salah satu metode perkembangan dari metode KNN, yang membedakan adalah terdapat kelas membership untuk menentukan kelas klasifikasi. Selain itu pada penelitian ini menggunakan kamus *Lexicon Based* untuk menentukan ekstraksi fitur kata. Hasil dari pengujian pada penelitian ini menunjukkan hasil akurasi nilai K=20 yang tertinggi sebesar 60% sedangkan nilai akurasi *precision* dan *recall* masing-masing mencapai 40% dan 40%. Pada pengujian *K-Fold Cross Validation* dengan *fold* sebesar 5 menghasilkan rata-rata sebesar 51,4%.

Kata Kunci : *tempat wisata, rating, klasifikasi, lexicon based dan fuzzy k-nearest neighbor*

Abstract

At present, with the ease of access to information, many tourist sites use rating features to help facilitate information. Rating is used as an indicator to support quality and popularity. Users only give an overall assessment of each comment and do not provide an assessment in accordance with the aspects discussed, making it difficult for comment readers to analyze the superior aspects of the comment. From this problem, in this study a rating classification system will be made on tourist attractions using the Fuzzy K-Nearest Neighbor (FKNN) method. FKNN method is one of the development methods of the KNN method, the difference is that there is a membership class to determine the classification class. In addition, this study uses a Lexicon Based dictionary to determine feature extraction. The results of the tests in this study showed the highest accuracy of K=20 values of 60% while the accuracy of precision and recall values reached 40% and 40% respectively. In testing the K-Fold Cross Validation with 5 fold it produces an average of 51.4%.

Keywords*: tourism, rating, classification, lexicon based and fuzzy k-nearest neighbor method*

1. PENDAHULUAN

Indonesia adalah negara kepulauan yang memiliki banyak keindahan. Keindahan itu menjadi daya tarik dan merupakan salah satu sumber penyokong sektor ekonomi penting di Indonesia yaitu sektor wisata. Sektor wisata tersebut haruslah didukung dengan kemudahan informasi untuk megakses segala sesuatu.

Pada masa sekarang ini, banyaknya website tujuan wisata menggunakan rating fitur untuk membantu mempermudah informasi. Rating digunakan sebagai indikator untuk menunjang kualitas dan popularitas (Rosi, et al., 2018). Komentar didapatkan dari wisatawan yang telah berkunjung ke tempat wisata tersebut. Dengan hanya melalui komentar, kita bisa dapat menentukan rating untuk tempat wisata. Salah satu website tentang tempat wisata adalah tripadvisor.com. Website ini berisi tentang rekomendasi tempat wisata beserta aksesnya. Adapun pengguna bisa memberikan komentar dan penilaian tempat wisata dengan menggunakan rating. Kekurangan dari sitem ini, rating atau penilaian

yang diberikan oleh beberapa pengguna tidak sesuai dengan aspek yang dibicarakan dalam komentar. sehingga menyulitkan pembaca komentar menganalisis aspek yang unggul pada komentar tersebut. Selain itu, rating yang ditulis oleh pengguna sangat bias serta ambigu.

Terkait masalah tersebut, penelitian ini bertujuan untuk mengelompokkan kata dari komentar yang diberikan pengguna sesuai dengan aspek yang terkait menjadi sebuah informasi dengan klasifikasi dalam bentuk rating. Metode klasifikasi merupakan metode yang memakai data *training* sebagai bagian dalam mengambil keputusan dan dapat menyesuaikan parameter-parameter yang dibutuhkan.

Sebelumnya, telah dilakukan penelitian serupa dalam menentukan klasifikasi tempat wisata. Adapun referensi penelitian terdahulu yang sudah dilakukan yaitu penelitian yang dilakukan oleh Ramadhani et al. (2018), Penelitian ini dilakukan di 30 destinasi wisata di Indonesia yang berfokus di pulau Jawa dan Bali. Proses menggunakan 3 jenis dataset. (Ramadhani, et al., 2017). Metode yang dipakai dalam penelitian ini adalah *Fuzzy K-Nearest Neighbor*. Metode ini digunakan karena dapat memberikan prediksi yang lebih akurat dibanding dengan metode K-nearest Neighbour. Analisis tersebut didapatkan dari penelitian oleh Baiq Findiarin Billyan dengan akurasi *Fuzzy K-Nearest Neighbor* dengan nilai k=1 dan k=2 nilainya mencapai 96%, k=3 hingga k=10 nilainya mencapai 76% (Billyan, et al., 2017).

Untuk mendapatkan hasil klasifikasi yang baik, diperlukan proses tambahan untuk mengekstrak kata. Dalam penelitian ini, fitur yang digunakan adalah kamus lexicon based. Kamus ini didapat dari penelitian oleh David Moeljadi dan Francis Bond tentang kamus WordNet dan SentiWord. WordNet adalah kamus yang berisi pangkalan data leksikal bahasa Melayu dan Indonesia. Sedangkan SentiWord adalah kamus yang berisi tentang sentimen kata dalam bahasa inggris (Moeljadi, et al., 2016).

Berdasarkan penjelasan diatas, penulis akan melakukan penelitian klasifikasi rating berdasarkan komentar tempat wisata di media sosial dengan menggunakan metode *Fuzzy K-Nearest Neighbor*.

2. DASAR TEORI

2.1 Tempat Wisata

Tempat wisata atau objek wisata adalah suatu wilayah yang mempunyai daya tarik untuk menyegarkan jasmani dan rohani. Semua tempat atau keadaan alam yang memiliki sumber daya wisata yang dikelola dan dikembangkan sehingga mempunyai daya tarik yang dikunjungi wisatawan. Tempat wisata ada berupa alam dan buatan. Daerah tempat wisata terdapat tiga hal yang menarik yakni adanya sesuatu yang dilihat untuk menarik wisatawan datang ke tempat tersebut. Kedua, adanya hal yang dapat dibeli di tempat tersebut serta adanya sesuatu yang dapat dilakukan yang mebuat pengunjung datang ke tempat wisata tersebut (Aprilia, et al., 2017).

2.2 *Preprocessing*

Preprocessing merupakan tahap awal dalam normalisasi data teks. Pada proses ini pengolahan data dapat mengubah data menjadi terstruktur sesuai kebutuhan untuk proses berikutnya. Data teks tersebut harus dipecah menjadi unsur kecil. Dokumen dipecah menjadi bab, bab dipecah menjadi sub-bab lalu dipecah menjadi paragraf, kalimat, kata dan yang terkecil menjadi suku kata.

Adapun beberapa tahap dalam proses ini meliputi *cleansing* yaitu penghapusan angka, karakter, serta hal lain selain huruf. Kemudian proses *case folding* yaitu mengubah semua huruf menjadi huruf kecil. Selanjutnya proses *tokenisasi* yaitu memisah kalimat menjadi kata dengan *delimiter* spasi. Proses selanjutnya *filtering* yaitu tahap mengambil kata valid dari hasil tokenisasi dengan menggunakan algoritme *stopword* (membuang kata tidak penting) atau *wordlist* (menyimpan kata penting). Proses terakhir adalah *Stemming* yaitu mengubah kata menjadi bentuk akarnya atau dasarnya. Pada proses ini menggunakan *stemming* Sastrawi dengan algoritma Nazief dan Adriani.

2.2. *Lexicon Based*

Dalam analisis sentimen dikenal adanya metode berbasiskan *Machine Learning*, berbasiskan kamus (*Lexicon Based*), atau gabungan keduanya (Nurfalah, et al., 2011). *Lexicon Based* memiliki beberapa alur yaitu pengambilan data dan membuat kamus. Pengambilan data diambil dari website atau media social berupa teks. Lalu pembuatan kamus yaitu proses normalisasi data dan ekstraksi kunci pada kata.

Kamus *Lexicon Based* yang digunakan dalam penelitian ini adalah kamus SentiWord dan WordNet yang berasal dari penelitian oleh David Moeljadi (Moeljadi, et al., 2016).

2.3. *K-Nearest Neighbor* (K-NN)

K-Nearest Neighbor (K-NN) merupakan metode klasifikasi pada data mining yang menggunakan algoritme *supervised*. Metode *Supervised learning* adalah metode untuk mendapatkan model atau aturan baru dengan cara mempertimbangkan model atau aturan data latih dengan data uji. *K-Nearest Neighbor* menggunakan klasifikasi ketetanggan untuk memprediksi data baru atau data testing. Terdapat beberapa cara mengukur tingkat kedekatan antara data testing dengan data training, yaitu dengan jarak *euclidean* atau jarak *manhattan*.

Proses yang dilakukan yaitu perhitungan jarak terdekat antara data latih dan data uji. Perhitungan ini bertujuan agar mengetahui jarak pada masing-masing *record*. Perhitungan ini menggunakan *euclidean distance* seperti pada Persamaan (1).

$$d_i = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2} \quad (1)$$

keterangan:

d_i = Jarak kedekatan
p = Jumlah atribut data
x_1 = Data latih
x_2 = Data uji

2.4. *Fuzzy K-Nearest Neighbor* (K-NN)

Fuzzy K-Nearest Neighbor merupakan metode gabungan antara *logika Fuzzy* dengan K-Nearest Neighbor. *Logika fuzzy* adalah suatu metode yang diterapkan pada masalah-masalah yang mengandung ketidakpastian. Penyelesaiannya dengan menggunakan himpunan pada masing-masing kelas yang penyajiannya dengan menggunakan fungsi karakteristik. Sedangkan *K-Nearest Neighbor* adalah metode klasifikasi dengan mencari nilai K terdekat dari tetangganya. Dalam metode ini, kedua metode tersebut digabung.

Sebelum menghitung nilai keanggotaan pada *Fuzzy K-Nearest Neighbor*, terlebih dahulu dilakukan proses menggunakan pada Persamaan (2).

$$u_{ij=}\begin{cases} 0{,}51 + \left(\frac{n_j}{n}\right) * 0{,}49, jika\ j = 1 \\ \left(\frac{n_j}{n}\right) * 0{,}49, jika\ j \neq 1 \end{cases} \quad (2)$$

keterangan :

n = Jumlah data latih
n_j = Jumlah anggota kelas j
j = Kelas data

Selanjutnya menghitung nilai keanggotaan masing-masing kelas dengan Persamaan (3).

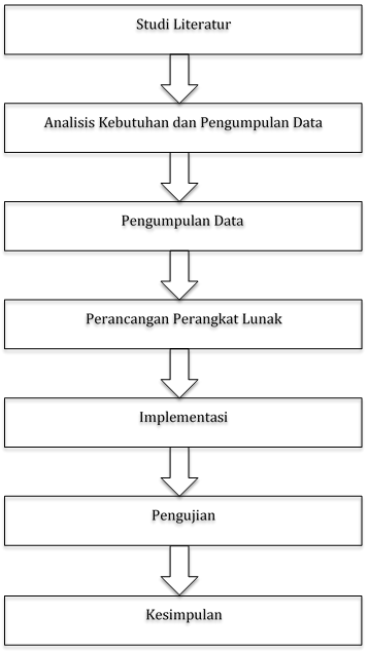
$$u(x, c_i) = \frac{\sum_{k=1}^{k} u(x_k, c_i) * d(x, x_k)^{\frac{-2}{(m-1)}}}{\sum_{k=1}^{k} d(x, x_k)^{\frac{-2}{(m-1)}}} \quad (3)$$

keterangan :

u_{ij} = nilai keanggotaan fuzzy
k = nilai tetangga terdekat
j = variabel data keanggotaan data uji
m = bobot yang besarnya m>1

3. METODOLOGI

Metodologi pada penelitian ini, menjabarkan mengenai alur yang digunakan dalam implementasi sistem. Adapun tahapan alur tersebut yakni mencari studi literatur, analisis kebutuhan dan pengumpulan data, perancangan sistem, implementasi, pengujian hingga pengambilan kesimpulan. Diagram alur metodologi dapat dilihat pada Gambar 1.



Gambar 1. Alur Metodologi peneltian

4. HASIL DAN PEMBAHASAN

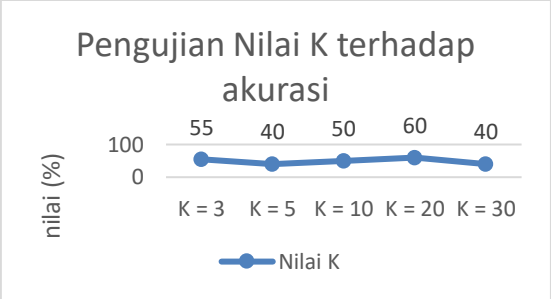
Pengujian pada penelitian ini, menggunakan data latih sebanyak 230 dan data uji sebanyak 48. Terdapat tiga pengujian pada penelitian ini yaitu:

4.1. Pengujian Pengaruh Nilai K Terhadap Akurasi

Pengujian ini dilakukan untuk mengetahui pengaruh banyak tetangga terhadap nilai akurasi yang dihasilkan. Selain itu untuk mengetahui nilai k yang paling optimal pada penentuan klasifikasi rating. Adapun nilai k yang digunakan bervariasi yaitu k = 3, k = 5, k = 10, k = 20 dan k = 30. Berikut hasil percobaan pengujian dapat dilihat pada Tabel 1 dan grafik pada Gambar 2.

Tabel 1. Pengaruh Nilai K terhadap Akurasi

Nilai K	Jumlah Data Latih	Akurasi
3	230	55%
5	230	40%
10	230	50%
20	230	60%
30	230	40 %



Gambar 2. Hasil pengaruh nilai K terhadap akurasi

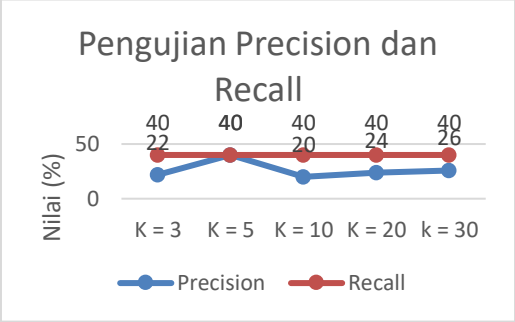
Pada Gambar 2 dapat diketahui akurasi dari pengujian pengaruh nilai K terbesar dihasilkan pada jumlah K=20 dengan akurasi sebesar 60%. Dari grafik diatas dapat dilihat bahwa nilai K yang berbeda menghasilkan akurasi yang berbeda. Sehingga dapat disimpulkan bahwa nilai K berpengaruh terhadap akurasi yang dihasilkan. Meskipun akurasi yang dihasilkan nilainya naik turun dan berubah tidak terlalu signifikan.

4.2. Pengujian *Precision* dan *Recall*

Pada pengujian *precision* dan *recall* data yang digunakan yaitu data latih sebesar 230 dan data uji sebesar 48. Hasil dari pengujian *precision* dan *recall* ditunjukkan pada Tabel 2 dan grafik pada Gambar 3.

Tabel 2. Pengujian Precision dan Recall

Nilai K	Jumlah Data Latih	*Precision*	*Recall*
3	230	22%	40%
5	230	40%	40%
10	230	20%	40%
20	230	24%	40%
30	230	26%	40%



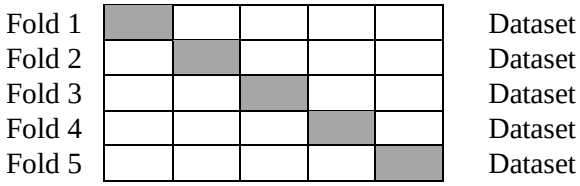
Gambar 3. Hasil pengujian precision dan recall

Analisis dalam pengujian *precision* dan *recall* adalah akurasi yang dihasilkan belum memuaskan. Hal tersebut dikarenakan penyebaran data yang tidak stabil sehingga per kelas data dapat mempengaruhi hasil pengujian.

Nilai pengujian *precision* mendapat akurasi yang sangat kecil dan juga naik turun. Hal tersebut dikarenakan banyak kelas data yang tidak bervariasi atau minim dan banyak data yang tidak mirip polanya. Sedangkan nilai *recall* cenderung stabil tetapi bernilai kecil. Beberapa data dalam dokumen mengandung *term* yang tidak termasuk ke dalam kategori yang sebenarnya dan jumlahnya banyak, dikarenakan hal tersebut banyak data yang tidak terklasifikasi dengan kategori yang sesuai yang dapat menurunkan nilai akurasi.

4.3. Pengujian *K-Fold Cross Validation*

Pada proses pengujian ini, dataset akan dibagi kedalam jumlah k subset dan selanjutnya dilakukan proses training dan testing sebanyak k kali. Proses ini sebagaimana ditunjukkan pada Gambar 4.

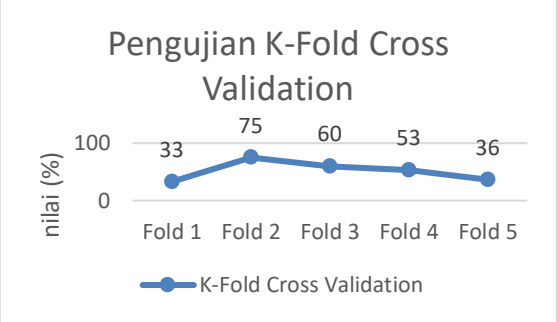


Gambar 4. Pembagian dataset *K-Fold Cross Validation*

Dalam pengujian ini menggunakan dataset berjumlah 150 dan *fold* berjumlah 5. Seluruh dataset dibagi sebanyak 5*fold* lalu tiap subset digunakan sebagai data testing dan sisanya digunakan sebagai data testing. Kemudian dilakukan sebanyak K kali. Berikut hasil pengujian *K-Fold Cross Validation* terdapat pada Tabel 3 dan grafik pada Gambar 5.

Tabel 3. Pengujian K-Fold Cross Validation

Fold	**Akurasi**
1	33%
2	75%
3	60%
4	53%
5	36%
Rata-Rata	51.4%



Gambar 5. Hasil pengujian *K-Fold Cross Validation*

Hasil pengujian ini mendapatkan hasil terbesar yaitu akurasi 75% pada nilai *fold* 2. Dari pengujian *K-Fold Cross Validation* mendapatkan rata-rata sebesar 51.4%.

Dari pemaparan hal tersebut, analisis pada pengujian ini dikatakan akurasi yang dihasilkan pada sistem ini belum maksimal. Analisis berikutnya yang di dapat dari pengujian ini adalah perbedaan nilai akurasi yang cukup jauh antara *fold* 1 dan *fold* 2 dikarenakan pada pengujian *fold* 1 data training yang digunakan cenderung memiliki karakteristik yang cukup berbeda dengan data testing, sedangkan pada *fold* 2 data training dan data testing cenderung memiliki karakteristik yang sama sehingga menghasilkan prediksi yang lebih baik.

5. PENUTUP

5.2. Kesimpulan

1. Metode *Fuzzy K-Nearest Neighbor* dalam menentukan klasifikasi rating pada tempat wisata pada penelitian ini dapat diterapkan. Sistem ini menggunakan lima kelas klasifikasi rating. Adapun proses utama yang dilakukan sebelum klasifikasi yaitu *preprocessing* dan penerapan kamus *Lexicon Based*. Pada *preprocessing* terdapat 5 tahap yaitu *cleansing*, *case folding*, tokenisasi, *filtering* dan *Stemming*. Setelah proses tersebut terdapat proses penarikan nilai sentimen dengan menggunakan kamus *Lexicon Based*. Selanjutnya pada proses klasifikasi dalam metode FK-NN yaitu, mencari nilai *Euclidean Distance*, dan proses *Fuzzy* untuk mencari nilai derajat keanggotaan pada masing-masing kelas sebagai penentu dalam prediksi.
2. Metode *Fuzzy K-Nearest Neighbor* menghasilkan akurasi yang kurang baik dalam menentukan klasifikasi rating tempat wisata berdasarkan beberapa mekanisme pengujian yang telah dilakukan yaitu:Pengujian pengaruh nilai K terhadap akurasi yang menghasilkan akurasi tertinggi pada K=20 yaitu sebesar 60%. Dengan menggunakan data latih 230 data. Berikutnya pengujian *precision* dan *recall* yang menghasilkan akurasi *precision* sebesar 40% dan *recall* sebesar40% pada jumlah data latih sebanyak 230 data latih. Terakhir pengujian *K-Fold Cross Validation* menghasilkan akurasi paling tinggi sebesar 75% pada *fold* 2 dikarenakan data yang karakteristik yang cukup mirip sedangkan yang paling rendah pada *fold* 1 sebesar 33%. Rata-rata hasil pengujian ini sebesar 51.4%.
3. Pada penelitian ini terdapat nilai K yang berbeda yang berpengaruh terhadap nilai akurasi yang dihasilkan. Meskipun akurasi yang berubah tidak terlalu signifikan. Hasil akurasi yang didapat nilainya naik turun, tetapi tidak terlalu jauh. Nilai akurasi yang dihasilkan pada K=3 sebesar 55%, K=5 sebesar 40%, K=10 sebesar 50%, K=20 sebesar 60%, dan K=30 sebesar 40%.

5.2. Saran

1. Untuk meningkatkan akurasi pada sistem, diharapkan pada penelitian selanjutnya menggunakan data latih yang lebih banyak sehingga semakin banyak karakteristik dari data latih maka akan meningkatkan kebenaran pada prediksi.
2. Akurasi yang dihasilkan penelitian ini kurang baik, diharapkan melakukan penelitian serupa dengan menggunakan metode klasifikasi lain atau menggunakan kombinasi metode yang berbeda.
3. Disarankan penelitian berikutnya melakukan *preprocessing* yang dapat menemukan adanya kata-kata ambigu dan tidak baku yang sulit ditemukan oleh kamus *Lexicon Based*.
4. Disarankan untuk membuat sistem yang bisa menyeleksi kata dengan makna sarkas yang dapat mempengaruhi hasil klasifikasi.
5. Dapat melakukan pemrosesan kata lebih dari satu suku kata yang dapat memaksimalkan hasil klasifikasi.

6. DAFTAR PUSTAKA

Agusta, L., 2009. Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief dan Adriani untuk Stemming Dokumen Teks Bahasa Indonesia. *Konferensi Nasional Sistem dan Informatika*, Bali.

Aprilia, E. R., Sunarti & Pangestuti, E., 2017. Pengaruh daya tarik wisata dan fasilitas layanan terhadap kepuasan wisatawan di pantai Balekambang Kabupaten Malang. *Jurnal Administrasi Bisnis (JAB)*, Vol. 51.

Billyan, B. F., Bhawiyuga, A. & Primananda, R., 2017. Implementasi Metode Klasifikasi Fuzzy K-Nearest Neighbor (FK-NN) Untuk Fingerprint Access Point Pada Indoor Positioning. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,*Vol. 1.

Jiang, S., Pang, G., Wu, M. & Kuang, L., 2011. An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications,* Volume 39, pp. 1503-1509.

Kurniawati, N. R., 2016. Penentuan destinasi wisata favorit berbasis aturan dan analisis sentimen pada tweet berbahasa Indonesia. *Electronic Theses and Desertation Universitas Gadjah Mada.*

Liu, B., Hu, M. & Cheng, J., 2015. Opinion observer: analysing and comparing opinions on the Web. *In Proceedings of the 14th international.*

Moeljadi, D. & Bond, F., 2016. Identifying and Exploiting Definitions in Wordnet Bahasa. Bucharest, *Proceedings of the Eighth Global WordNet Conference*.

Moeljadi, D., Le, T. A., Miura, Y. & Ohkuma, T., 2016. Sentiment Analysis for Low Resource Languages: *Proceedings of the 12th Workshop on Asian Language Resources.*

Ramadhani, D. M., Rahmat, C. & Rahutomo, F., 2017. Tourism Destination Rating System Based on Social Media Analysis. *International Conference on Sustainable Information Engineering and Technology (SIET)*.

Shofa, R. A., Muflikhah, L. & Ridok, A., 2016. Penerapan Metode Fuzzy K-Nearest Neighbor (FK-NN) untuk Menentukan Kualitas Hasil Rendemen Tanaman Tebu. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK).*

Tala, F. Z., 2003. A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. *M.S. thesis. M.Sc. Thesis. Master of Logic Project. Institute for Logic, Language and Computation. Universiteti van Amsterdam The Netherland.*

Wilianto, L., Pudjiantoro, T. H. & Umbara, F. R., 2017. Analisis Sentimen terhadap Tempat Wisata dari Komentar Pengunjung. *Prosiding of SNATIF*.