Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.202184508
Vol. 8, No. 2, April 2021, hlm. 389-394 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

ANALISIS SENTIMEN MENGENAI MODA RAYA TERPADU (MRT) JAKARTA DENGAN METODE BM25 DAN K-NEAREST NEIGHBOR

Indriati*[1], Bayu Rahayudi[2], Candra Dewi[3]

[1,2,3]Program Studi Teknik Infomatika, Fakultas Ilmu Komputer Universitas Brawijaya
Email: [1]indriati.tif@ub.ac.id, [2]ubay1@ub.ac.id, [3]dewi_candra@ub.ac.id
*Penulis Korespondensi

(Naskah masuk: 16 Desember 2020, diterima untuk diterbitkan: 22 Maret 2021)

Abstrak

Moda Raya Terpadu (MRT) Jakarta merupakan alat transportasi berkecepatan tinggi berupa kereta rel listrik yang ada di ibukota Jakarta . Adanya banyak tanggapan positif maupun negatif dari masyarakat dapat dipergunakan sebagai masukan bagi operator layanan MRT Jakarta untuk terus bisa memperbaiki pelayanan demi terwujudnya angkutan massal yang berguna bagi masyarakat. Proses pengumpulan data tanggapan dapat diperoleh dari sosial media maupun komentar-komentar di setiap pemberitaan mengenai MRT Jakarta. Data-Data tersebut akan diolah dulu dengan melewati tahapan preprocessing untuk diklasifikasikan menjadi sentimen yang bersifat positif maupun sentimen yang bersifat negatif. Metode untuk mengklasifikasikan tanggapan adalah K-Nearest Neighbor dengan menggunakan metode BM25 sebagai metode untuk mengetahui kesamaan antar data. Proses pengujian yang digunakan pada penelitian ini yaitu cross validation dengan *k-fold* sebanyak 5. Pengujian dilakukan dengan jumlah data data uji sebanyak 130 dokumen dan data latih sebanyak 520 dokumen untuk setiap fold. Berdasarkan rata-rata hasil pengujian diperoleh hasil terbaik pada nilai k=11 dengan nilai *f-measure* sebesar 0,89088, *recall* sebesar 0,934286, dan *precision* sebesar 0,852351. Hasil pengujian menunjukkan nilai k yang semakin besar menghasilkan nilai *f-measure* yang semakin kecil karena proses klasifikasi berjalan kurang baik dengan banyaknya tetangga yang tidak sama kelasnya dengan data uji digunakan untuk menentukan kelas data uji.

Kata kunci: analisis sentimen, MRT Jakarta*, BM25, K-Nearest Neighbor*

SENTIMENT ANALYSIS FOR MASS RAPID TRANSPORTATION (MRT) IN JAKARTA USING BM25 AND K-NEAREST NEIGHBOR METHOD

Abstract

Mass Rapid Transportation (MRT) is a rapid transit transportation system using electric rail trains in Jakarta. This kind of transportation is one of Jakarta Provincial Government Programs so that people would switch to public transport. Many positive and negative comments from the public can be used as input for MRT service operators to improve services for the realization of mass transportation that is useful for the community. Therefore, needed a fast process for processing the comments that can be classified as positive sentiment and negative sentiment. The data collection process for comment can be obtained from social media and comments on any news regarding MRT Jakarta.These data will be processed first by passing the preprocessing stage and then classified into positive sentiments and negative sentiments. The method used to classify is K-Nearest Neighbor and using BM25 as a method to determine the similarity between data. The testing process used in this research is cross validation with 5-folds. Each test uses 130 documents for test documents and 520 documents for training data. The average results obtained in each test process produce the best results at a value of k = 11 with an f-measure of 0.89088, a recall of 0.934286, and a precision of 0.852351. The test results show that the greater k value results in a smaller f-measure value because the classification process is not going well with the number of neighbors who are not the same class as the test data used to determine the class of the test data.

Keywords: *sentiment analysis, BM25, K-Nearest Neighbor*

1. PENDAHULUAN

Moda Raya Terpadu (MRT) Jakarta merupakan alat transportasi berkecepatan tinggi berupa kereta rel listrik yang ada di ibukota Jakarta. Layanan MRT ini diberi nama "Ratangga". PT. MRT Jakarta sebagai operator layanan ini merupakan badan usaha milik daerah yang berada

dibawah kepemilikan Pemerintah Provinsi DKI Jakarta (Wikipedia). Jenis transportasi ini merupakan salah satu cara Pemerintah Provinsi DKI Jakarta agar masyarakat mau beralih ke transportasi umum. Adanya transportasi ini diharapkan dapat mengurangi tingkat kemacetan yang tiap hari selalu terjadi di DKI Jakarta.

MRT yang beroperasi resmi sejak 24 Maret 2019 menimbulkan banyak tanggapan dari masyarakat. Adanya banyak tanggapan positif maupun negatif dari masyarakat dapat dipergunakan sebagai masukan bagi operator layanan MRT Jakarta untuk terus bisa memperbaiki pelayanan demi terwujudnya angkutan massal yang berguna bagi masyarakat. Tanggapan yang negatif bisa dijadikan referensi perbaikan dari segi pelayanan maupun sarana yang akan dilakukan operator layanan MRT Jakarta.

Tanggapan masyarakat mengenai MRT Jakarta bisa disampaikan lewat sosial media maupun tanggapan-tanggapan yang ada di kolom pemberitaan mengenai MRT Jakarta. Tanggapan yang sangat banyak jumlahnya menyebabkan sulitnya menganalisis sentimen positif atau negatif yang ada di masyarakat. Analisis sentimen merupakan ilmu agar suatu barang, layanan, organisasi, individu, masalah dan peristiwa dapat diketahui penilaiannya dari pengguna atau orang lain berdasarkan pendapat, sentimen, evaluasi, penilaian, sikap dan emosi (Liu,2012). Penilaian tersebut dapat dibagi menjadi tiga yaitu penilaian yang sifatnya positif, negatif, dan netral.

Analisis sentimen merupakan sebuah penelitian mengenai pendapat, emosi serta sentimen yang diungkapkan melalui tulisan (Liu, 2010). Dari beberapa pengertian sentimen analisis dapat diambil kesimpulan bahwa sebuah dokumen, kalimat, atau bentuk tulisan yang lain dapat diklasifikasikan oleh analisis sentimen menjadi kelas sentimen positif , netral dan negatif. Akan tetapi lebih banyak penelitian mengenai analisis sentimen membagi menjadi dua kelas saja yaitu positif dan negatif karena dianggap kelas netral tidak diperlukan.

Ada berbagai algoritma yang dapat digunakan untuk mengklasifikasikan data yang berupa tulisan. Salah satu algoritma atau metode yang dapat digunakan adalah *K-Nearest Neighbor* (KNN). KNN adalah algoritma yang menggunakan data-data dengan jarak yang artinya data-data tersebut mempunyai kesamaan fitur dengan data yg lain untuk diklasifikasikan menjadi satu kelas (Prasetyo,2012). Dalam penelitian (Nugraha, Faraby, & Adiwijaya, 2015) juga dijelaskan bahwa algoritma KNN adalah algoritma pengklasifikasian dengan menggunakan jarak terdekat antara 2 buah objek. Algoritma KNN juga mempunyai tujuan agar dapat mengidentifikasi objek berdasarkan atribut dan sampel data latih sehingga untuk proses klasifikasinya suatu titik query akan menemukan objek k dari data latih dengan jarak k yang terdekat dengan titik query (Hardiyanto, Rahutomo, & Puspitasari, 2016).

Proses algoritma KNN untuk melakukan klasifikasi suatu data yaitu dengan menentukan jumlah k tetangga terdekat untuk diambil kelas mayoritas yang menjadi penentu kelas data tersebut. Kinerja dan hasil yang akurat merupakan salah satu keunggulan algoritma KNN apabila digunakan untuk klasifikasi teks (Suharno, Fauzi dan Perdana, 2017). K-Nearest Neighbor dapat menghasilkan nilai akurasi sebesar 90% dengan nilai k sebesar 5 sehingga K-Nearest Neighbor dapat dikatakan memiliki kinerja yang baik (Bagaskoro, Fauzi dan Adikara, 2018).

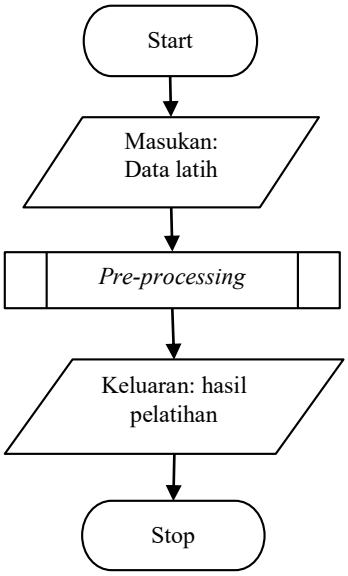
Proses pembobotan dan pemeringkatan dokumen merupakan salah satu bagian penting dalam klasifikasi teks. Salah satu metode yang dapat digunakan untuk pemeringkatan dokumen adalah BM25. BM25 adalah salah satu metode pemeringkatan yang dapat mengurutkan hasil kecocokan antara kata kunci dengan dokumen-dokumen yang ada di korpus atau koleksi data. Selain BM25 ada juga beberapa metode lain yang dapat digunakan misalnya TF-IDF. Akan tetapi menurut penelitian (Yang dkk., 2012), pembobotan TF-IDF yang dibandingkan dengan BM25 memperoleh hasil yang lebih rendah. Sehingga dapat dikatakan kinerja BM25 lebih baik daripada pembobotan TF-IDF. Penelitian tersebut menggunakan data bug report sehingga dapat mendeteksi adanya duplikasi pada bug report dengan akurasi sebesar 90%. Selain itu penelitian (Tinega, et al., 2018) juga memperkuat bahwa pemeringkatan menggunakan BM25 menghasilkan nilai yang jauh lebih baik apabila dibandingankan Boolean Model dan Vector Space Model. Penelitian (Whissel & Clarke, 2013) juga menunjukkan apabila dibandingkan dengan cosinus similarity maka algoritma BM25 merupakan algoritma yang hasil pemeringkatannya lebih baik daripada menggunakan cosinus similarity

Hasil akurasi dari penggabungan metode BM25 dengan KNN memiliki hasil akurasi terbaik yaitu 88,97%. Nilai tersebut diperoleh dengan mengunakan lima kelas dan nilai k sebesar 10 untuk klasifikasi status merokok seorang pasien (Aramaki dkk., 2006). Dari penelitian-penelitian yang telah diuraikan maka dapat diambil kesimpulan bahwa BM25 dan KNN dapat digunakan untuk melakukan analisis sentimen pada tanggapan-tanggapan yang diberikan masyarakat mengenai MRT Jakarta. Penelitian ini lebih fokus penerapan metode BM25 pada proses pembobotan dan pemeringkatannya.

2. METODE PENELITIAN

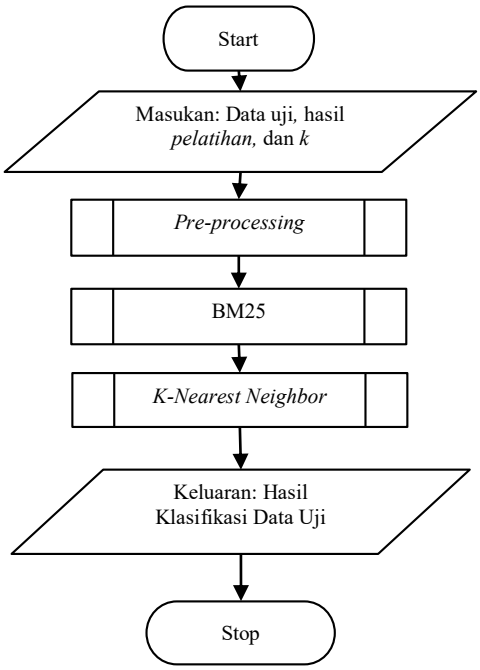
Data yang berupa tanggapan masyarakat terhadap MRT Jakarta diperoleh dari komentar-komentar yang ada dalam berita online mengenai MRT Jakarta. Data yang dipergunakan pada penelitian ini berjumlah 650 data tanggapan dengan rincian data uji sebanyak 130 dan data latih 520.

Pelabelan dilakukan dengan cara voting dari 3 orang responden sehingga misal terhadap satu dokumen Responden 1 dan Responden 2 memberikan label positif dan Responden 3 memberikan label negatif maka dokumen tersebut untuk pengujiannya mendapatkan label positif. Gambaran secara umum mengenai algoritma yang digunakan pada penelitian ini dapat dilihat pada Gambar 1 dan Gambar 2. Pada Gambar 1 menjelaskan alur yang dilakukan untuk melakukan proses pada data latih sebelum dilakukan proses pembobotan BM25 dan klasifikasi KNN. Sedangkan pada Gambar 2 menjelaskan alur yang dilalui mulai dari mulai dari proses preprocessing data uji hingga melakukan proses klasifikasi dengan memanfaatkan nilai-nilai yang sudah didapatkan terlebih dahulu melalui proses-proses yang ada pada Gambar 1. Proses pada Gambar 2 melibatkan data uji yang nantinya akan dihitung nilai kesamaannya dengan menggunakan metode BM25 dengan setiap data latih.



Gambar 1. Alur Proses Pelatihan

Gambar 1 menjelaskan dokumen yang menjadi data latih akan diproses pada tahapan *pre-processing* teks sehinga teks yang awalnya bentuknya tidak terstruktur menjadi teks yang bentuknya terstruktur. Tahap *pre-processing* terdiri dari *case folding* untuk mengubah seluruh huruf pada data latih menjadi huruf kecil, *tokenization* untuk memisahkan kalimat menjadi kata-kata, *stopword removal* untuk menghilangkan kata-kata yang tidak penting, dan *stemming* untuk mengubah kata berimbuhan menajdi kata dasar. Kemudian setiap kata hasil *pre-processing* teks akan dilakukan perhitungan frekuensi kemudian disimpan untuk proses pengujian pada Gambar 2. Data uji, hasil pelatihan dan nilai k akan menjadi input untuk proses pengujian. Pada proses ini data uji juga akan melalui proses pre processing yang sama persis dengan proses pelatihan. Hasil preprocessing data uji dan hasil pelatihan akan digunakan untuk proses perhitungan BM25 yang terdiri dari proses perhitungan frekuensi tiap kata dan nilai IDF. Apabila nilai tersebut sudah diperoleh maka dapat dilanjutkan proses berikutnya yaitu menghitung BM25 *score* untuk setiap dokumen pada data latih terhadap data uji yang digunakan pada proses pengujian. Setelah mendapatkan nilai BM25 *score* tiap data latih terhadap data uji maka tahapan selanjutnya yaitu dilakukan proses klasifikasi dengan menggunakan algoritma KNN. Tahapan yang dilakukan pada KNN yaitu mengurutkan nilai BM25 score mulai dari nilai yang terbesar hingga terkecil kemudian memilih sejumlah nilai *k* yang ditentukan pada input proses pengujian. Penentuan kelas data uji berdasarkan pada kelas mayoritas dari data latih yang mempunyai nilai kesamaan terbesar sejumlah k yang sudah ditentukan sebagai parameter algoritma KNN.



Gambar 2. Alur Proses Pengujian

3. TEXT MINING

Text mining merupakan salah satu cara untuk mengetahui pola yang ada dalam kumpulan teks sebagai penerapan konsep penambangan data. Selain itu *text mining* dapat didefinisikan sebagai penambangan dari data yang berupa teks. Pola-pola data, tren serta mencari inti dari sebuah informasi merupakan manfaat yang diperoleh dari text mining.

Proses *text mining* dibagi menjadi empat tahapan yaitu teks (*text preprocessing*) sebagai proses yang pertama kali dilakukan, tahapan selanjutnya adalah transformasi teks (*text transformation*) untuk mengubah teks kedalam bentuk yang dapat dikomputasi, setelah itu dilakukan tahapan pemilihan fitur-fitur yang sesuai (*feature selection*) dan tahapan yang paling akhir yaitu penemuan pola (*pattern discovery)* (Feldman dan Sanger, 2006).

3.1 *Preprocessing*

Text pre-processing merupakan proses yang pertama kali dilakukan untuk mempersiapkan *dataset* agar pemrosesan data teks dapat dilakukan dengan lebih mudah selain itu agar dapat diperoleh kinerja yang tinggi (Feldman dan Sanger, 2006). Perubahan data teks dari tidak terstruktur menjadi terstruktur merupakan manfaat dari Text Preprocessing.C*ase folding*, *tokenization*, *stopword removal*, dan *stemming* merupakan proses-proses yang dilakukan pada text preprocessing.

3.1.1. Case Folding

Case folding adalah proses yang dilakukan untuk mengubah semua huruf pada data teks menjadi huruf besar untuk semua teks atau huruf kecil untuk semua teks (Feldman dan Sanger, 2006).

3.1.2. Tokenization

Tokenization adalah proses yang dilakukan untuk memisahkan data teks yang berupa bab, subbab, paragraf maupun kalimat menjadi kata-kata. (Feldman dan Sanger, 2006).

3.1.3. Stopword Removal

Stopword removal adalah proses yang dilakukan untuk menghilangkan kata-kata yang tidak penting agar dapat meningkatkan kecepatan pemrosesan teks serta memperkecil memori yang digunakan untuk menyimpan kata-kata yang digunakan sebagai fitur (Feldman dan Sanger, 2006).

3.1.4. Stemming

Stemming adalah proses untuk mengubah kata yang berimbuhan menjadi dasar. Pada teks berbahasa Indonesia, proses stemming ini dilakukan dengan cara mengenali dan menghilangkan *suffix*, *prefix*, dan *confix* agar kata kembali ke bentuk kata dasar (Adriani dkk., 2007).

3.2 BM25

Proses pembobotan dan pemeringkatan dokumen merupakan salah satu tahapan dalam text mining. Tahapan tersebut dapat dilakukan dengan menggunakan berbagai macam metode. Salah satu metode yang dapat digunakan ada metode BM25. Metode BM25 merupakan metode untuk melakukan pemeringkatan agar diperoleh urutan hasil dokumen berdasarkan kesamaan kata kunci yang dicari terhadap dokumen-dokumen yang ada pada koleksi dokumen. Persamaan (1) merupakan persamaan yang digunakan untuk menghitung nilai kesamaan menggunakan metode BM25.

$$BM25 = \sum_{i=1}^{|q|} idf(q_i).\frac{tf(q_i,d)\,.\,(k_1+1)}{tf(q_i,d)+k_i\,.\,(1-b+b\,.\,\frac{|d|}{dl_{avg}})} \quad (1)$$

Keterangan:
$idf(qi)$: nilai *invers document frequency* pada *term query ke i*
$tf(q_i, d)$: nilai jumlah frekuensi *term query* ke *i* pada dokumen ke *j*
k_1 : $1{,}2 \leq k_1 \leq 2{,}0$
b : $0{,}5 \leq b \leq 0{,}8$
dl_{avg} : nilai rata-rata panjang semua dokumen
$|d|$: nilai panjang dokumen

Perhitungan nilai *idf* yang ada pada Persamaan (1) ditunjukkan pada Persamaan (2).

$$idf(q_i) = \log(\frac{N - df(q_i) + 0{,}5}{df(q_i) + 0{,}5}) \quad (2)$$

Keterangan:
$idf(qi)$: nilai *invers document frequency* pada *term query i*
N : nilai total dokumen dalam koleksi
$df(qi)$: nilai jumlah dokumen yang terdapat *term query i*

3.3 K-Nearest Neighbor

K-Nearest Neighbor (KNN) merupakan algoritma klasifikasi yang dapat digunakan untuk mengklasifikasikan data yang berupa teks atau tulisan. Kesamaan antara satu dokumen dengan dokumen lainnya yang menjadi dasar perhitungan algoritma *K-Nearest Neighbor*. Nilai kesamaan antar dokumen tersebut dihitung dengan menggunakan metode BM25. Secara garis besar proses perhitungan KNN dengan menggunakan nilai BM25 dapat dijelaskan sebagai berikut:

1. Melakukan perhitungan nilai BM25 score antara data uji dengan seluruh data latih.
2. Melakukan pengurutan nilai BM25 *score* terhadap data latih sebanyak nilai k yang ditentukan pada algoritma KNN.
3. Penentuan kelas dari data uji berdasarkan kelas mayoritas yang muncul pada k data latih yang terpilih pada proses no.2

4. HASIL PENGUJIAN DAN ANALISIS

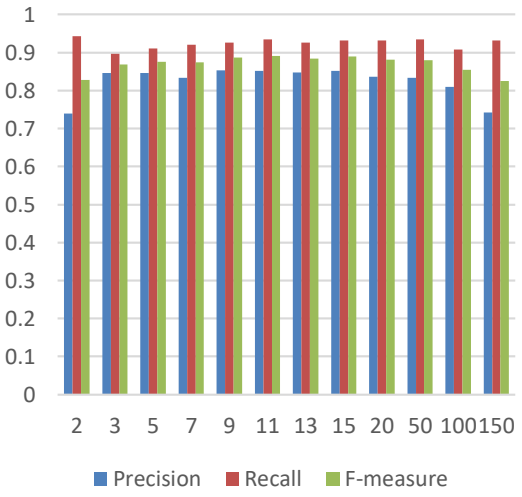
4.1. Pengaruh Nilai *K* dari Hasil *Precision, Recall,* dan *F-Measure*

Proses pengujian dilakukan untuk memperoleh nilai *precision, recall,* dan *f-measure* sebagai evaluasi terhadap sistem klasifikasi dokumen. Pada proses pengujian ini untuk memastikan setiap data pernah menjadi data latih dan data uji maka dilakukan cross validation dengan pengujian sebanyak 5 fold. Setiap pengujian menggunakan data uji sebanyak 130 dokumen dengan setiap foldnya menggunakan dokumen yang berbeda-beda. Pengujian yang dilakukan sebanyak 5 kali untuk setiap nilai k sehingga nilai yang diperoleh dari rata-rata perhitungan *precision, recall* dan *f-measure* yang

ditampilkan pada Tabel 1 dan digambarkan pada Gambar 3.

Tabel 1. Rata – Rata *Precision, Recall,* dan *F-Measure*

k	*Precision*	*Recall*	*F-measure*
2	0,739845	0,942857	0,82819
3	0,845787	0,897143	0,8693
5	0,845685	0,911429	0,87613
7	0,834225	0,92	0,87477
9	0,852897	0,925714	0,88747
11	**0,852351**	**0,934286**	**0,89088**
13	0,846912	0,925714	0,884
15	0,852108	0,931429	0,88922
20	0,837072	0,931429	0,88103
50	0,833957	0,934286	0,8805
100	0,810266	0,908571	0,85525
150	0,742168	0,931429	0,82471



Gambar 3. Rata – Rata *Precision*, *Recall*, dan *F-Measure*

Seperti yang terlihat pada Gambar 3, dengan menggunakan nilai k yang berbeda-beda maka nilai evaluasi yang terdiri dari precision, recall dan f-measure yang dihasilkan juga berbeda sehingga dapat disimpulkan bahwa nilai k memengaruhi hasil klasifikasi untuk algoritma KNN. Gambar 3 juga memperlihatkan hasil yang berfluktuasi untuk rentang nilai *k*=1 sampai dengan *k*=20. Sedangkan pada saat rentang nilai *k*=50 sampai dengan *k*=150, nilai *precision*, *recall*, dan *f-measure* akan mengalami penurunan.

Adanya perbedaan jumlah kelas negatif dengan jumlah kelas positif mengakibatkan proses klasifikasi untuk data uji dengan label kelas negatif mengalami salah klasifikasi. Hal ini terjadi dikarenakan jumlah data kelas negatif lebih sedikit daripada data kelas positif sehingga apabila penentuan kelas berdasarkan kelas mayoritas pada data k tetangga maka kemungkinan data uji menjadi kelas positif lebih besar dibandingkan kelas negatif. Nilai evaluasi terbaik didapatkan ketika nilai *k*=11 dengan nilai *f-measure* sebesar 0,89088. Pada nilai *k*=11 metode BM25 dapat memeringkatkan dokumen dari yang sama kelasnya dengan data uji lebih banyak dibandingkan yang berbeda kelasnya dengan data ujai, sedangkan nilai k yang semakin besar akan membuat nilai precisionnya semakin kecil karena banyak dokumen yang berbeda kelasnya dengan data uji akan ikut dihitung untuk penentuan kelas data uji meskipun dengan banyaknya dokumen yang menjadi tetangga terdekat akan memperbesar nilai recall.

Penelitian ini menunjukkan bahwa metode BM25 dapat melakukan pemeringkatan dokumen yang sesuai kelasnya dengan data uji , sedangkan untuk algoritma *K-Nearest Neighbor* akan memperoleh hasil yang lebih baik jika data latih yang dipergunakan jumlahnya seimbang karena sangat di pengaruhi oleh nilai k.

5. KESIMPULAN

Berdasarkan pengujian yang telah dilakukan dan analisis terhadap pengujian, dapat ditarik kesimpulan bahwa Metode BM25 dan K-Nearest Neighbor dapat dimanfaatkan untuk melakukan klasifikasi dokumen komentar mengenai MRT Jakarta. Tahapan yang harus dijalankan agar memperoleh hasil klasifikasi yaitu tahapan pertama adalah *pre-processing,* dilanjutkan tahapan kedua yaitu perhitungan BM25, dan tahapan terakhir untuk mengklasifikasikan dokumen dengan algoritma *K-Nearest Neighbor.* Pada saat nilai k sebesar 11 akan menghasilkan nilai terbaik untuk f- measure yaitu 0,89088. Jumlah nilai *k* memengaruhi hasil klasifikasi pada penelitian ini. Semakin kecil nilai *k* yang digunakan, sistem berjalan dengan baik, sedangkan semakin besar nilai *k* yang digunakan sistem berjalan kurang baik.

DAFTAR PUSTAKA

ARAMAKI, E., IMAI, T., MIYO, K. DAN OHE, K., 2006. Patient Status Classification by using Rule based Sentence Extraction and BM25-kNN based Classifier. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. [daring] Tersedia pada: <http://luululu.com/paper/2006-i2b2/i2b2smoking.pdf>.

BAGASKORO, G.N., FAUZI, M.A. DAN ADIKARA, P.P., 2018. Penerapan Klasifikasi Tweets Pada Berita Twitter Menggunakan Metode K-Nearest Neighbor Dan Query Expansion Berbasis Distributional Semantic. 2(10), hal.3849–3855.

BING LIU. 2010, "Sentiment Analysis and Subjectivity", Handbook of natural language processing 2: 627-666.

CHAOVALIT, P., ZHOU, L., 2005, "Moview review mining: A comparison between supervised and unsupervised classification approaches", in

Proceedings of the Hawaii International Conference on System Sciences.

HARDIYANTO, E., RAHUTOMO, F., & PUSPITASARI, D. (2016). Implementasi K Nearest Neighbor (KNN) pada Klasifikasi Artikel Wikipedia Indonesia.

LAROSE, DANIEL T. 2005, "Discovering Knowledge in Data: An Introduction to Data Mining".

LIU, B. 2012. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.

NUGRAHA, P., FARABY, A. A., & ADIWIJAYA. (2015). Klasifikasi Dokumen Menggunakan Metode k-Nearest Neighbor (kNN) dengan Information Gain. Seminar Nasional Teknologi Informasi dan Multimedia, ISSN : 2302 - 3805, 5(1), hal.1–14.

PRASETYO, EKO. 2012, "Data Mining: Konsep dan Aplikasi menggunakan Matlab", Andi

RUSSELL, S.J., NORVIG, P., 2010. Artificial Intelligence: A Modern Approach Third Edition. Upper Saddle River: New Jersey.

SUHARNO, C.F., FAUZI, M.A. DAN PERDANA, R.S., 2017. Klasifikasi Teks Bahasa Indonesia pada Dokumen Pengaduan Sambat Online Menggunakan Metode K-Nearest Neighbors dan Chi-Square. Systemic: Information System and Informatics Journal, 3(1), hal.25–32.

TINEGA, G. A., MWANGI, W. & RIMIRU, R., 2018. Text Mining in Digital Libraries using OKAPI BM25 Model. International Journal of Computer Applications Technology and Research, 7(10), pp. 398-406.

WANG, L. & ZHAO, X., 2012. Improved KNN Classification Algorithms Research in Text Categorization.

WHISSEL, J. S. & CLARKE, C. L., 2013. Effective Measures for Inter-Document Similarity. Cikm, p. pp.1361–1370.

YANG, C., DU, H., WU, S. DAN CHEN, I., 2012. Duplication Detection for Software Bug Reports based on BM25 Term Weighting.