Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 4, No. 11, November 2020, hlm. 4103-4109 http://j-ptiik.ub.ac.id

Analisis Sentimen Ulasan Produk Kecantikan Menggunakan Metode BM25 dan *Improved K-Nearest Neighbor* dengan Seleksi Fitur *Chi-Square*

Dewi Syafira[1], Indriati[2], Sutrisno[3]

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

Abstrak

Pengaruh produk kecantikan merupakan hal yang mulai diminati oleh kaum perempuan. Dengan kemudahan yang diberikan saat ini terdapat platform khusus berbasis web maupun mobile phone yaitu *Female Daily. Female daily* merupakan situs media informasi yang berisi konten produk kecantikan tentang perawatan wajah hingga tubuh untuk siap diberi ulasan oleh konsumen yang telah mencoba atau sedang menggunakan produknya. Data ulasan dapat digunakan sebagai acuan sebelum konsumen ingin mencoba produk kecantikan. Banyaknya ulasan mengakibatkan konsumen sulit untuk memilih produk yang di inginkan. Pada penelitian ini membantu konsumen untuk mengetahui data ulasan tesebut masuk kedalam sentimen positif atau sentimen negatif. Proses dalam Analisis Sentimen memerlukan metode BM25 yang digunakan sebagai pembobotan kata, *Improved K-Nearest Neighbor* sebagai penentuan dalam memilah sentimen dan Seleksi Fitur *Chi-Square* untuk mengurangi jumlah kata dalam klasifikasi pada teks. Pengujian dilakukan menggunakan *5-fold cross validation* dengan hasil terbaik diperoleh pada nilai k= 15 menghasilkan rata-rata nilai sebesar presisi= 0,9, *recall*= 0,8, *accuracy*= 0,7806 dan *f-measure*= 0,8428 selanjutnya dari pengujian seleksi fitur *Chi-Square* berdasarkan persentase dengan parameter k=15 didapatkan hasil tertinggi pada persentase sebanyak 40% dan 50% dengan nilai presisi= 0,888, *recall*= 0,8, *accuracy*= 0,7818 dan *f-measure*= 0,842.

Kata kunci: *analisis sentimen, Female Daily, BM25, Improved K-Nearest Neighbor, Chi-Square*

Abstract

The impact of beauty product which makes women interest. With the convenience provided at this time there is a special platform based on the web and mobile phone that is Female Daily. Female Daily is an information media site that contains the content of beauty products about facial care to the body to be given reviews by consumers who have tried or using the product. Review Data can be used as a reference before consumers want to try out beauty products. A lot of reviews resulted makes consumers has difficult to choose the product they want. In this research, it helps consumers to know the review data is entered into positive or negative sentiments. The process in sentiment analysis requires the BM25 method used as a weighted word, Improved K-Nearest Neighbor as a determination in sorting the sentiments and Chi-Square feature selection to reduce the number of words in the classification on the text. The test used is 5-fold cross validation with the best results on k-rank=15 resulting on the average value of precision= 0.9, recall= 0.8, accuracy = 0.7806 and f-measure= 0.8428. Then, testing used selection of Chi-square features selection based on percentages with k-rank= 15 parameters, the highest percentage value of 40% and 50% with the value of precision = 0.888, recall = 0.8, accuracy = 0.7818 and f-measure = 0.842.

Keywords: *sentiment analys, Female Daily, BM25, Improved K-Nearest Neighbor, Chi-Square*

1. PENDAHULUAN

Pengaruh produk kecantikan merupakan hal yang mulai diminati oleh masyarakat saat ini terutama kaum perempuan yang ingin merawat tubuh dari kepala hingga kaki. Dengan kemudahan yang diberikan saat ini yaitu mulai banyak pengembangan-pengembangan aplikasi mengenai ulasan-ulasan tentang produk kecantikan di mana memiliki manfaat untuk memudahkan dalam menemukan produk yang ingin dicari dan dibutuhkan. *Female Daily.com*

merupakan platfrom khusus berbasis web maupun *mobile phone* yang berisi konten produk kecantikan tentang perawatan wajah hingga tubuh untuk siap diberi ulasan oleh konsumen yang telah mencoba atau sedang menggunakan produknya. Tak hanya itu saja dalam aplikasi *Female Daily* terdapat artikel seputar pengetahuan mengenai produk yang memiliki bahan-bahan yang terkandung di dalamnya sehingga dapat membandingkan zat-zat apa saja yang akan dipakai dalam suatu produk tersebut.

Dengan adanya ulasan-ulasan yang diberikan pada aplikasi *Female Daily* maka analisis sentimen memiliki pengaruh dan manfaat yang sangat besar. Analisis sentimen merupakan bidang studi untuk menganalisis opini, sentimen, evaluasi, penilaian, sikap seseorang dan emosi terhadap suatu produk, layanan, organisasi, individu, masalah, peristiwa dan lainnya, sehingga dapat di klasifikasikan menjadi dua kategori yaitu positif dan negatif (Liu, 2012).

Penggunaan analisis sentimen digunakan untuk ulasan pada aplikasi *Female Daily* dengan tujuan dapat mengetahui kualitas dari suatu produk apakah produk tersebut bagus untuk digunakan atau tidak direkomendasi. Penerapan penelitian menggunakan metode BM25 dan metode *Improved K-Nearest Neighbor* dengan Seleksi Fitur Chi Square. Terdapat macam-macam seleksi fitur yang sering ditemukan dalam penelitian-penelitian sebelumnya dengan menggunakan seleksi fitur *Information Gain* dan *Chi-Square*.

Perbedaan yang didapatkan dari penelitian sebelumnya bahwa pada kategorisasi klasifikasi teks bahasa indonesia diketahui jika semakin banyak jumlah dokumen yang digunakan dalam klasifikasi dapat meningkatkan nilai *F-Measure* klasifikasi teks dengan metode *Chi-Square*, sedangkan dengan metode information gain dapat menurunkan nilai *F-Measure* klasifikasi teks (Sofiana, et al., 2012). *Chi-Square* dapat menghilangkan fitur penggangu yang ada dalam proses klasifikasi dokumen dengan tujuan hasil yang diperoleh menghasilkan peningkatan nilai akurasi yang maksimal dalam klasifikasi teks dibandingkan tidak menggunakan seleksi fitur.

Setelah mendapatkan fitur-fitur yang telah diseleksi selanjutnya tahap pemeringkatan dengan menggunakan metode BM25. Metode BM25 ialah metode yang menghasilkan keluaran berdasarkan pemeringkatan dari hasil perhitungan pembobotan kata tiap dokumen yang lebih baik daripada menggunakan *cosinus similarity* (Whissel & Clarke, 2013) dan juga metode *Improved K-Nearest Neighbor* merupakan tahap selanjutnya dari metode BM25 ialah *Improved K-nearest Neighbor* merupakan suatu algoritme modifikasi dari metode *K-Nearest Neighbor* dengan menggunakan penetapan nilai K yang berbeda pada tiap kategori menyesuaikan jumlah data latih di mana dengan menggunakan *Improved K-Nearest Neighbor* mendapatkan hasil akhir lebih stabil dibandingkan menggunakan *K-Nearest Neighbor* (Wang & Zhao, 2012).

Pada penelitian sebelumnya yang dilakukan pada tahun 2019 oleh Ardhimas Ilham Bagus Pranata yang meneliti dengan judul Klasifikasi Dokumen pada Laporan Kepolisian dengan menggunakan metode BM25 dan *Improved K-Nearest* mendapat hasil penelitian menggunakan pemeringkatan BM25 dan metode *Improved K-Nearest Neighbor* yang dapat digunakan dalam melakukan klasifikasi pada dokumen laporan kepolisian menghasilkan nilai rata-rata tertinggi *precision*= 0,953373, *recall*= 0,931382, *f-measure*= 0,938122, serta *accuracy*= 0,956795 (Pranata, et al., 2019).

Oleh karena itu, dari permasalahan yang telah dipaparkan oleh penulis, penulis melakukan penelitian dengan menerapkan Analisis Sentimen Ulasan Produk Kecantikan menggunakan metode BM25 dan *Improved K-Nearest Neighbor* dengan Seleksi Fitur *Chi-Square.* Penelitian ini diharapkan dapat membantu dalam menganalisis menemukan kelebihan atau kekurangan serta mengevaluasi dari suatu produk sehingga mudah dalam mencari produk yang diinginkan sesuai dengan kebutuhan konsumen

2. DASAR TEORI

2.1 Female Daily

Female daily ialah suatu aplikasi platform khusus berbasis web maupun *mobile phone* yang berisi konten produk kecantikan tentang perawatan wajah hingga tubuh untuk siap diberi ulasan oleh konsumen (Pujadayanti, et al., 2018).

2.2 Analisis Sentimen

Analisis sentimen merupakan bidang studi untuk menganalisis opini, sentimen, evaluasi, penilaian, sikap seseorang dan emosi terhadap suatu produk, layanan, organisasi, individu, masalah, peristiwa dan lainnya, sehingga dapat di klasifikasikan menjadi dua kategori yaitu positif dan negatif (Liu, 2012). Demikian juga

dengan penelitian sebelumnya menurut (Cvijkj & Michahelles, 2011) analisis sentimen digunakan untuk memahami ulasan atau komentar yang diciptakan oleh pengguna dan menjelaskan bagaimana sebuah produk maupun kualitas merek diterima oleh pengguna.

2.3 Preprocessing Text

Preprocessing diperlukan untuk memilih kata yang akan digunakan sebagai indeks. Indeks berupa kata-kata yang mewakili dokumen. *Preprocessing Text* merupakan suatu proses pengubahan bentuk data yang belum terstruktur menjadi data yang terstruktur sesuai dengan kebutuhan.

2.3.1 *Case Folding*

Case folding merupakan proses tahapan di mana dalam tiap dokumen semua huruf akan diubah menjadi huruf kecil.

2.3.2 *Tokenizing*

Tokenizing atau tokenisasi adalah proses memecah suatu teks menjadi kata, frasa, simbol atau elemennya yang memiliki makna berdasarkan tiap kata penyusunnya. Proses ini dilakukan untuk penghilangan angka, tanda baca dan karakter selain huruf alfabet, dan juga singkatan akronim yang harus diubah menjadi bentuk kata dasar (Kannan & Gurusamy, 2014)

2.3.3 *Filtering*

Filtering merupakan tahap pemilihan kata-kata penting yang di ambil dari hasil tokenisasi dengan menghapus kata yang tidak memiliki makna dari tiap dokumen.

2.3.4 *Stemming*

Stemming ialah proses pengubahan bentuk kata-kata berimbuhan menjadi kata dasar atau tahap untuk mencari root kata dari tiap kata hasil *filtering* (Kannan & Gurusamy, 2014).

2.4 Seleksi Fitur *Chi-Square*

Seleksi Fitur merupakan menghilangkan sejumlah fitur-fitur yang tidak ada hubungannya dengan kategori pada dokumen sehingga tujuannya ialah memilih fitur-fitur yang penting dan relevan pada dokumen untuk proses kategorisasi (Sofiana, et al., 2012). Proses perhitungan seleksi fitur *Chi-Square* dapat dilihat pada Persamaan 1.

$$\boldsymbol{X^2(t,c)} = \frac{\boldsymbol{N(AD-CB)^2}}{\boldsymbol{(A+C)(B+D)(A+B)(C+D)}} \quad (1)$$

Keterangan:

t = kata
c = kategori/kelas
N = banyaknya data latih
A = banyaknya dokumen yang memiliki term t pada kelas c
B = banyaknya dokumen yang memiliki term t selain pada kelas c
C = banyaknya dokumen yang tidak memiliki term t pada kelas c
D = banyaknya dokumen yang tidak memiliki term t selain pada kelas c

2.5 BM25

Metode BM25 merupakan metode *Best Match* 25 yang digunakan untuk menampilkan dokumen yang relevan dengan menggunakan mekanisme pemeringkatan berdasarkan tingkat kemiripan isi dari query yang diberikan sebanyak data latih (Pranata, et al., 2019). BM25 adalah fungsi peringkat yang dibuat oleh Stephen Robertson dan Karen Sparck Jones. Pada dasarnya metode BM25 memiliki 3 faktor dalam mempengaruhi pembobotan kata yaitu frequensi kemunculan query yang terdapat dalam dokumen yang dikenal juga sebagai TF(*Term Frequency*). Kemudian tahap kedua ialah IDF (*Inverse Document Frequency*) yang merupakan nilai invers dari total dokumen yang memiliki kata yang dicari. Selanjutnya tahap ketiga yaitu nilai dari rata-rata panjang dokumen (Russel & Norvig, 2010). Berikut merupakan perhitungan nilai pemeringkatan BM25 pada Persamaan 2.5.

$$Score(D,Q) = \sum_{i=1}^{N} IDF(q_i) \frac{TF(q_i,d_j).(k+1)}{TF(q_i,d_j)+k(1-b+b\frac{|d_j|}{L})} \quad (2)$$

Serta fungsi yang digunakan untuk menghitung nilai IDF dapat dilihat pada Persamaan 3.

$$IDF(q_i) = \log \frac{N-DF(q_i)+0{,}5}{DF(q_i)+0{,}5} \quad (3)$$

Keterangan:

$Score\ (D,Q)$ = Score kemiripan query dengan dokumen
$IDF(q_i)$ = Inverse Document Frequency q_i
$TF(q_i,d_j)$ = *Term Frequency* q_i pada dokumen d_j
L = Nilai rata-rata jumlah dokumen
N = Jumlah dokumen
$|d_j|$ = banyaknya kata dalam dokumen d_j
$DF(q_i)$ = Jumlah dokumen yang mengandung term q_i

k = 2
b = 0,75

2.6 Improved K-Nearest Neighbor

Improved K-nearest Neighbor yaitu algoritme yang merupakan modifikasi dari metode *K-Nearest Neighbor*. Kelebihan pada Algoritme ini ialah modifikasi yang dihasilkan pada *Improved K-Nearest Neighbor* terletak pada jumlah nilai k. Perbedaan dari *K-Nearest Neighbor* dengan *Improved K-nearest Neighbor* ialah jika pada *K-Nearest Neighbor* memiliki penetapan nilai k yang sama pada tiap masing-masing kategori sedangkan untuk *Improved K-Nearest Neighbor* memiliki penetapan jumlah nilai k yang berbeda untuk setiap kategori (Baoli, et al., 2003). Perhitungan untuk penetapan nilai k baru dapat dilihat pada Persamaan 4.

$$n = \left[\frac{k_1 * N(C_m)}{Maks\{n(C_m)|j=1...N_C}\right] \quad (4)$$

Serta menghitung probabilitas dari dokumen uji dengan menggunakan Persamaan 5, Persamaan 6 dan Persamaan 7.

$$W1 = \sum d_j \in topnkNN\ (c_m) sim(x, d_j) y(d_j, c_m) \quad (5)$$

$$W2 = \sum d_j \in top\ n\ k\ NN\ (c_m) sim(x, d_j) \quad (6)$$

$$p(x, c_m) = argmax_m \frac{W1}{W2} \quad (7)$$

Keterangan :

$p(x, c_m)$ = Peluang dokumen x anggota dari c_m

$\sum d_j \in top\ n\ k\ NN$ = Sejumlah nilai n tetangga tertinggi dari data latih

$sim(x, d_j)$ = Nilai similaritas dokumen x dengan dokumen latih d_j

$y(d_j, c_m)$ = Fungsi atribut 1 apabilla d_j masuk dalam c_m , 0 jika tidak

2.7 Evaluasi

Dalam penelitian ini pengujian menggunakan *Confusion Matrix* untuk mengukur evaluasi sistem. Terdapat pengukuran evaluasi sistem dengan kriteria diantaranya yaitu: *f-measure*, *precission*, *recall*, serta *accuracy* (Powers, 2011)

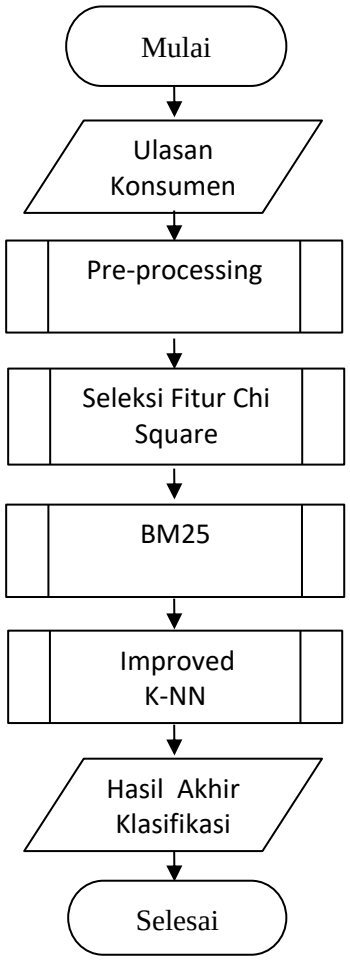
3. METODE PENELITIAN

3.1 Data Penelitian

Pengumpulan data yang digunakan untuk penelitian ini adalah data primer. Data primer merupakan data yang diambil langsung dari ulasan produk kecantikan berbahasa indonesia pada aplikasi *Female Daily* dan dapat diakses juga melalui link *https://femaledaily.com/*. Data yang digunakan sebanyak 150 data dan telah dilabeli oleh seorang pakar.

3.2 Perancangan Sistem

Proses yang dilakukan untuk perancangan sistem terdiri dari *preprocessing*, seleksi fitur *Chi-Square* yang digunakan untuk menyeleksi fitur kata, BM25 untuk menghitung pembobotan kata dan *Improved K-Nearest Neighbor* untuk hasil klasifikasi. Berikut rancangan algoritme keseluruhan pada Gambar 1.



Gambar 1 Alur Algoritme

4. HASIL PENGUJIAN DAN ANALISIS

Pengujian dilakukan dengan menggunakan *5-Fold Cross Validation* dan pengujian seleksi fitur *Chi-Square* berdasarkan jumlah persentase.

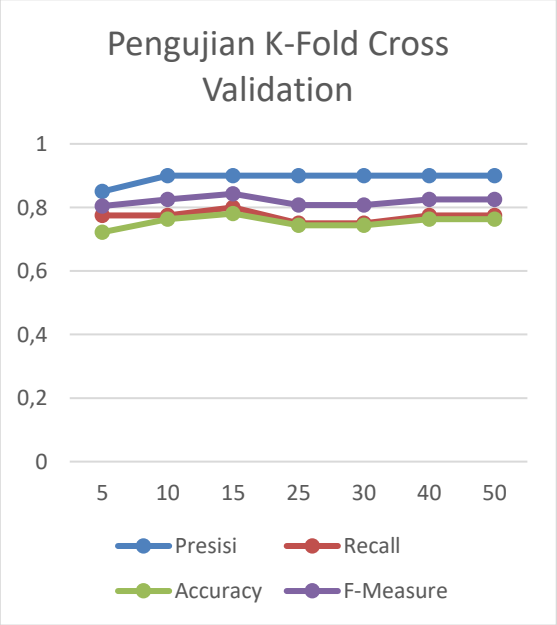
4.1. Pengujian dan Analisis *K-Fold Cross*

Validation

Pengujian ini memiliki fungsi untuk mengetahui perubahan dari hasil dari data uji terhadap data latih menjadi optimal. Hal ini dapat menghasilkan perhitungan nilai rata-rata presisi, *recall*, *accuracy* dan *f-measure* dengan parameter nilai k awal dari *Improved K-Nearest Neighbor* sebesar kelipatan 5 dan 10 dan parameter persentase *Chi-Square* yang diambil terbaik sebesar 50%. Hasil pengujian nilai rata-rata *K-Fold Cross Validation* dapat ditunjukkan pada Tabel 1.

Tabel 1 Hasil Pengujian K-fold Cross Validation

K	Presisi	Recall	Accuracy	F-measure
5	0,85	0,775	0,722	0,8044
10	0,9	0,775	0,763	0,8252
15	0,9	0,8	0,7806	0,8428
25	0,9	0,75	0,7438	0,8076
30	0,9	0,75	0,7438	0,8076
40	09	0,775	0,763	0,8252
50	09	0,775	0,763	0,8252



Gambar 2 Grafik rata-rata pengujian

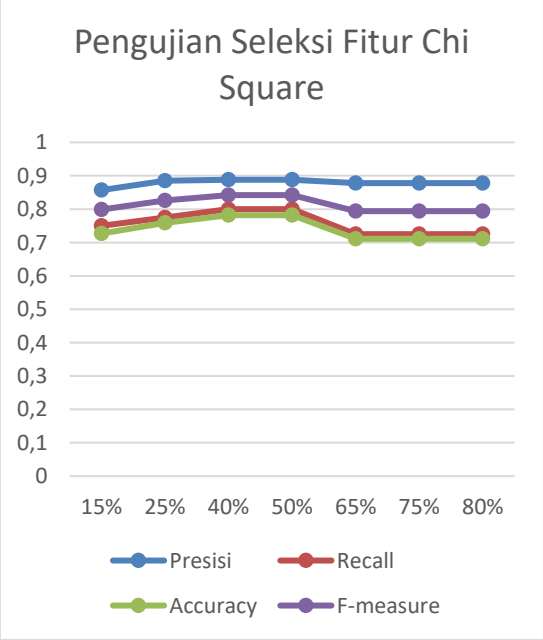
Dapat dilihat pada gambar 2 menghasilkan grafik hasil pengujian *K-fold Cross Validation* mengalami penurunan hingga peningkatan. Terdapat kata yang sering muncul dari hasil seleksi fitur *Chi-Square* seperti "muncul", "harga", "jerawat", "kantong", "cocok" dan sebagainya. Diketahui hasil rata-rata presisi, *recall*, *accuracy* dan *f-measure* tertinggi terdapat pada nilai awal k=15 yaitu nilai presisi= 0.9, *recall*= 0.8, *accuracy*= 0.7806 dan *f-measure*=0.8428 yang dipengaruhi dengan nilai k baru dari *Improved K-Nearest Neighbor* sebesar 15 untuk kelas positif dan 12,2 untuk kelas negatif. Dan hasil rata-rata terendah pada jumlah nilai k=25 dan k=30 yang dipengaruhi dengan nilai awal k=25 untuk nilai k baru dari *Improved K-Nearest Neighbor* sebesar 25 untuk kelas positif dan 20,45 untuk kelas negatif. Pengaruh nilai k awal juga mengalami perubahan pada nilai k baru *dari Improved K-Nearest Neighbor* yang menghasilkan nilai probabilitas sehingga nilai k baru dari *Improved K-Nearest Neighbor* akan mempengaruhi nilai dari ketetanggaan terdekat dengan data uji yang diberikan. Ini membuktikan bahwa jumlah k tidak berpengaruh untuk meningkatkan hasil evaluasi. Dikarenakan pada nilai k=40 hingga nilai k=50 justru mengalami penaikan dengan rata-rata nilai presisi = 0.9, *recall* = 0.775, *accuracy* = 0.763, dan *f-measure* = 0.8252.

4.2. Pengujian dan Analisis Seleksi Fitur *Chi-Square* Berdasarkan Jumlah Persentase

Pengujian ini dilakukan untuk mengetahui hasil presisi, *recall*, *accuracy serta f-measure* berdasarkan jumlah persentase pada seleksi fitur *Chi-Square* dengan menggunakan parameter nilai K= 15 serta jumah persentase sebanyak 15%, 25%, 40%, 50%, 65%, 75%, dan 80%. Pada Tabel 2 merupakan tabel hasil pengujian Seleksi Fitur *Chi-Square* berdasarkan jumlah persentase.

Tabel 2 Hasil Pengujian *Chi-Square*

Persentase	Presisi	*Recall*	*Accuracy*	*F-measure*
15%	0,857	0,75	0,727	0,799
25%	0,885	0,775	0,759	0,826
40$	0,888	0,8	0,7818	0,842
50%	0,888	0,8	0,7818	0,842
65%	0,878	0,725	0,711	0,794
75%	0,878	0,725	0,711	0,794
80%	0,878	0,725	0,711	0,794



Gambar 3 Grafik Hasil Pengujian *Chi-Square*

Dapat dilihat pada gambar 3 bahwa hasil grafik tertinggi terdapat pada persentase 40% dan persentase 50% yaitu presisi = 0.888, *recall* = 0.8, *accuracy* = 0,7818 dan *f-measure*= 0,842. Jumlah persentase pada hasil sebesar 40% dan 50% mengalami hasil yang sama dikarenakan fitur dari kata yang digunakan tidak jauh berbeda dengan hasil yang diperoleh sebelumnya dari jumlah persentase 40%. Namun dalam persentase 65% hingga 80% mengalami penurunan. Hal ini disebabkan fitur kata yang diambil merupakan fitur kata dengan perolehan nilai tertinggi dari hasil seleksi fitur *Chi-Square* sehingga sangat berpengaruh dalam klasifikasi. Semakin besar nilai persentase yang diambil tidak menjamin hasil evaluasi menjadi lebih baik. Sedangkan tanpa menggunakan seleksi fitur *Chi- Square* hasil akurasi yang didapatkan sebesar 0,6326. Dengan demikian, dapat disimpulkan bahwa dengan menggunakan seleksi fitur *Chi-Square* memiliki peran penting dalam proses klasifikasi dengan evaluasi yang lebih baik dibandingkan tanpa menggunakan seleksi fitur *Chi-Square*.

5. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan maka dapat disimpulkan bahwa Metode BM25 mampu menghasilkan fungsi sebagai pembobotan kata, *Improved K-Nearest Neighbor* sebagai penentuan untuk klasifikasi dari kelas positif dan kelas negatif, dan seleksi fitur *Chi Square* berfungsi menyeleksi kata yang sering muncul akan diambil sehingga kata yang tidak terlalu penting akan diabaikan atau dihilangkan. Serta dengan menggunakan metode BM25 dan *Improved K-Nearest Neighbor* dengan seleksi fitur *Chi-square* menghasilkan nilai rata-rata tertinggi terdapat pada nilai k awal yaitu k= 15 dengan parameter jumlah persentase seleksi fitur *Chi-Square* sebanyak 50% yaitu nilai presisi= 0.9, *recall*= 0.8, *accuracy*= 0.7806 dan *f-measure*=0.8428 serta menghasilkan nilai rata-rata terendah pada nilai k awal yaitu nilai k=25 dan k=30.

6. DAFTAR PUSTAKA

Baoli, L., Shiwen, Y. & Qin, L., 2003. An Improved K-Nearest Neighbor Algorithm for Text Categorization.

Cvijkj, I. P. & Michahelles, F., 2011. Understanding Social Media Marketing: A Case Study on Topics, Categories and Sentiment on a Facebook Brand Page.

Kannan, D. S. & Gurusamy, V., 2014. Preprocessing Techniques for Text Mining.

Liu, B., 2012. Sentiment Analysis and Opinion Mining. *Morgan And Claypool Publisher*, p. 7.

Powers, D., 2011. Evaluation: From Precision, Recall AND F-measure To ROC,. *Journal of Machine Learning Technologies*, pp. 37-38.

Pranata, A. I. B., Indriati & Marji, 2019. Klasifikasi Dokumen pada Laporan Kepolisian dengan Menggunakan Metode BM25 dan Improved K-Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, p. 2.

Pujadayanti, I., Fauzi, M. A. & Sari, Y. A., 2018. Prediksi Rating Otomatis pada Ulasan Produk Kecantikan dengan Metode Naïve Bayes dan N-gram. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, p. 4422.

Russel, S. & Norvig, P., 2010. *Artificial Intelligence A Modern Approach*. 3nd Edition ed. Upper Saddle River, New Jersey: Pearson Education.

Sofiana, I., Atastina, I. & Suryani, A. A., 2012. Analisis Pengaruh Feature Selection Menggunakan Information Gain dan Chi-Square Untuk Kategorisasi Teks Bahasa Indonesia.

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*, pp.1361–1370.