Terakreditasi SINTA Peringkat 2
Surat Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti No. 10/E/KPT/2019
masa berlaku mulai Vol. 1 No. 1 tahun 2017 s.d Vol. 5 No. 3 tahun 2021

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 3 No. 3 (2019) 505 – 510 ISSN Media Elektronik: 2580-0760

Seleksi Fitur Berbasis Pearson Correlation Untuk Optimasi Opinion Mining Review Pelanggan

Nova Tri Romadloni[1], Hilman F Pardede[2]
[1,2]Magister Ilmu Komputer, Fakultas Ilmu Komputer, STMIK Nusa Mandiri Jakarta
[2]Pusat Penelitian Informatika LIPI, Bandung
[1]novatrir2501@nusamandiri.ac.id, [2]hilman@nusamandiri.ac.id

Abstract

The comments contained on e-commerce users generally contain opinions about positive or negative experiences at several online shops. Sentences that can be written indirectly both a little or a lot, will affect other potential customers. So as a result of these comments cause a product sold at an online store has a rating of two things namely "recommended" or "non-recommended". However, detection of positive and negative opinions manually will require more time because of the large amount of data. For this reason opinion mining using technology in data mining can be used to automate positive and negative detection of comments. However, one of the main problems in opinion mining is limited data but has a large number of attributes. In this study, we propose the application of Pearson correlation (PC) based feature selection for opinion mining optimization. The results of the experiment show that the application of PC increases the performance of opinion mining systems in 3 types of classification, namely Logistic Regression, Naïve Bayes and Support Vector Machine, resulting in more optimal accuracy, namely 98.80%, 87.87% and 98.12%.

Keywords: *Pearson Correlation, Logistic Regression, Naïve Bayes, Support Vector Machine, Opinion Mining.*

Abstrak

Komentar yang terdapat pada pengguna *e-commerce* umumnya berisi pendapat mengenai pengalaman positif atau negatif pada saat beberlanja pada toko online. Kalimat yang dituliskan dapat secara tidak langsung baik sedikit maupun banyak, akan berpengaruh pada calon pelanggan yang lain. Sehingga akibat dari komentar tersebut menyebabkan suatu produk yang dijual pada toko online memiliki penilaian dua hal yaitu "*recommended*" atau "*non-recommended*". Namun, deteksi opini positif dan negatif secara manual akan membutuhkan waktu yang lebih lama karena banyaknya data. Untuk itu opinion mining menggunakan teknologi pada data mining dapat digunakan untuk otomatisasi deteksi positif dan negatif suatu komentar. Akan tetapi salah satu masalah utama pada opinion mining adalah data yang terbatas namun memiliki jumlah attribute yang besar. Pada penelitian ini, kami mengusulkan penerapan seleksi fitur berbasis pearson correlation (PC) untuk optimasi opinion mining. Hasil percobaan menunjukan penerapan PC meningkatkan kinerja sistem opinion mining pada 3 jenis klasifikasi yaitu Logistic Regression, Naïve Bayes dan Support Vector Machine menghasilkan akurasi lebih optimal yaitu 98,80%, 87,87 % dan 98,12%.

Kata kunci: *Pearson Correlation, Logistic Regression, Naïve Bayes, Support Vector Machine, Opinion Mining.*

1. Pendahuluan

Perkembangan ekonomi digital saat ini membawa banyak perubahan pada masyarakat salah satunya adalah perubahan perilaku konsumen yang menyukai berbelanja secara online. Belanja online menjadi tren baru yang terasa lebih sederhana, efisien dam cepat tanpa menemui hambatan. Proses ini menjadi pilihan yang banyak disukai oleh beberapa kalangan masyarakat untuk berbelanja atau membeli beberapa produk sesuai dengan kebutuhan. Hal tersebut termasuk dalam sebuah pilihan yang mayoritas mengganggap banyak kelebihan atau banyak manfaat yang didapatkan dari hasil transaksi tersebut. Diantaranya seperti menghemat biaya transportasi dan waktu berbelanja terasa akan sangat menjadi lebih hemat dan efektif [1].

Ulasan online yang diberikan konsumen kepada penjual akan mempengaruhi penilaian atau rating penjual pada platfrom ecommerce yang bersangkutan.

Diterima Redaksi : 21-08-2019 | Selesai Revisi : 01-12-2019 | Diterbitkan Online : 11-12-2019

[2]. Informasi dalam ulasan produk atau penilaian pengguna dapat memberikan dampak positif dan negatif pada pihak perusahaan. Ulasan produk atau penilaian yang diberikan oleh pengguna pada platform *ecommerce* juga dapat menjadi informasi bagi pengguna lainnya. *E-WOM* dalam bentuk ulasan produk atau penilaian pengguna dalam platform digital tidak hanya berfungsi sebagai informasi untuk pengguna lain, tetapi juga sebagai *recommender* [3].

Dalam mendekteksi komentar jika dilakukan secara manual akan membutuhkan waktu yang banyak karena harus diperiksa satu persatu [4]. Data yang digunakan cukup banyak dan sulit dilakukan karena bentuk dari komentar yang tidak beraturan serta membutukan biaya yang mahal karena melalui tahapan atau serangkaian proses manual yang menyesuaikan banyaknya data. Maka dibutuhkan sebuah pendekatan untuk membantu mengetahui hasil dari *opinion mining*.

Pendekatan menggunakan *machine learning* untuk melakukan *opinion mining* memiliki kelebihan dan kekurangan. Salah satunya adalah akurasi dari pendekatan klasifikasi *machine learning* sangat baik, akan tetapi peforma klasifikasinya domain dependent terhadap dataset yang digunakan pada saat training [5]. Sedangkan untuk proses klasifikasi pada teks menggunakan banyak fitur sehingga diperlukan seleksi fitur untuk mengurangi dimensi fitur dan mendapatkan kombinasi fitur yang optimal [6]. Seleksi fitur merupakan bagian penting untuk mengoptimalkan kinerja dari metode klasifikasi. Tujuan utama dari seleksi fitur adalah mengurangi kompleksitas, meningkatkan akurasi dan memilih fitur optimal dari suatu kumpulan fitur data [7].

Pada penelitian ini menggunakan seleksi fitur yang berbasis pearson correlation dengan algoritma *Logistric Regression, Naïve Bayes* dan *Support Vector Machine* sehingga didapatkan nilai akurasi yang lebih baik.

Identifikasi masalah yang dilakukan anatra lain:

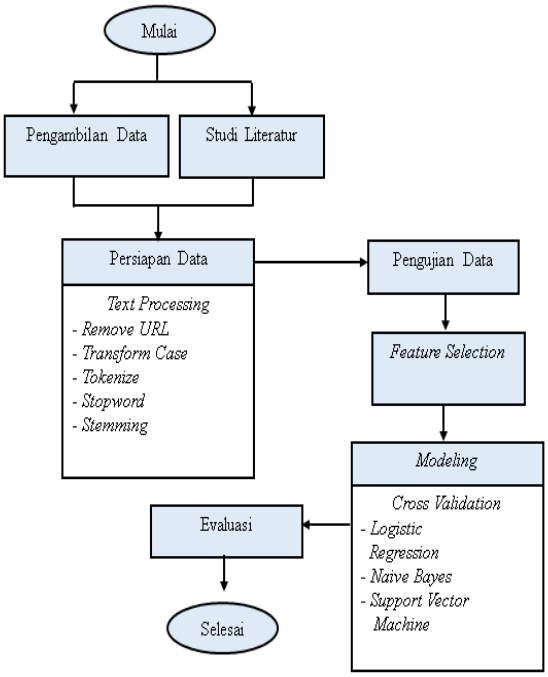
1. Bagaimana cara mengoptimalkan *opinion mining review* pelanggan dengan menggunakan seleksi fitur *pearson correlation* ?
2. Bagaimana hasil penerapan algoritma *Logistic Regression, Naïve Bayes* dan *Support Vector Machine* dalam *opinion mining review pelanggan* ?

2. Metode Penelitian

Penelitian yang membahas mengenai langkah-langkah dalam penelitian untuk menunjang kegiatan dalam penelitian. Tahapan penelitian yang dilakukan, disusun dan dijelaskan secara berurutan.

2.1. *Preprocessing Data*

Proses persiapan data atau *data preparation* merupakan hal yang penting untuk tahap selanjutnya, yaitu mengurangi atribut yang kurang berpengaruh terhadap proses klasifikasi data yang dimasukan pada tahap ini masih berupa data mentah yang masih kotor, sehingga hasil proses ini adalah dokumen berkualitas yang harapannya mempermudah dalam proses klassifikasi. Pada tahap ini dimulai dengan membagi yang didapat, data dibagi dalam label dan atribut yang telah ditentukan dan membuang data-data yang tidak dibutuhan dalam keperluan analisa, sehingga didapat hasil data yang siap dianalisa.



Gambar 1. Tahapan Penelitian

Text mining dapat didefinisikan secara luas sebagai suatu proses menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan tools analisis yang merupakan komponen-komponen dalam data mining yang salah satunya adalah kategorisasi. Proses ini terdiri dari beberapa tahap pembersihan dokumen seperti berikut.

Tahapan *Tokenizing* adalah proses memecah dokumen menjadi kumpulan kata. *Tokenization* dapat dilakukan dengan menghilangkan tanda baca dan memisahkannya per-spasi. Tahapan ini juga menghilangkan karakter-karakter tertentu seperti tanda baca dan mengubah semua token ke bentuk huruf kecil (*lower case*). Tahapan *Filtering* atau *Stopword Removal* adalah Stopwords removal merupakan proses penghilangan kata tidak penting pada deskripsi melalui pengecekan kata-kata hasil parsing deskripsi apakah termasuk di dalam daftar kata tidak penting (*stoplist*). Tahapan *Stemming* adalah Stemming merupakan bagian yang tidak terpisahkan dalam Information Retrieval (IR). Tidak banyak algoritma yang dikhususkan untuk stemming bahasa Indonesia dengan berbagai keterbatasan didalamnya. Algoritma Porter misalnya, algoritma ini membutuhkan waktu yang relatif lebih singkat dibandingkan dengan stemming

menggunakan algoritma Nazief dan Adriani, namun proses stemming menggunakan algoritma Porter memiliki persentase keakuratan lebih kecil dibandingkan dengan stemming menggunakan algoritma Nazief dan Adriani. Algoritma Nazief dan Adriani sebagai algoritma stemming untuk teks berbahasa Indonesia yang memiliki kemampuan persentase keakuratan lebih baik dari algoritma lainnya.

2.2. Seleksi Fitur dengan Pearson Correlation

Analisis korelasi digunakan untuk mengetahui kekuatan antara hubungan korelasi kedua variabel dimana variabel lainnya dianggap berpengaruh dikendalikan atau dibuat tetap (sebagai variabel control). Karena variabel yang diteliti adalah data interval maka menggunakan teknik *pearson correlation* [8].

Nilai yang dihasilkan pada Pearson's terletak pada [-1;1], untuk nilai -1 yang berarti korelasi negatif sempurna (karena satu variabel meningkat, yang lainnya menurun), +1 berarti korelasi positif sempurna dan 0 yang berarti tidak ada korelasi linier antara kedua variabel tersebut [9].

Nilai Pearson's untuk mendapatkan fitur – fitur dan dilakukan proses seleksi fitur berdasarkan nilai Pearson's tertinggi dan mendapatkan fitur – fitur yang terseleksi berdasarkan nilai Pearson's.Penelitian yang dilakukan oleh Matthew Shardlow, menjelaskan bahwa seleksi fitur dengan menggunakan metode filter, Pearson Correlation Coefficient dapat meningkatkan akurasi dan memiliki kelebihan yaitu efektif dan cepat untuk jumlah fitur yang sangat banyak [10].

2.3. *Modelling*

Sebagai *Classifier,* kami menggunakan 3 jenis *Classifier* untuk menguji efektifitas metode kami, yaitu: Logistik Regresi adalah metode yang paling umum digunakan dalam pendekatan untuk membuat model prediksi probabilitas kejadian suatu peristiwa seperti halnya regresi linear. Logistic regression ini hanya digunakan jika variabel ouput dari model yang digunakan didefinisikan sebagai katagori biner. Perbedaan metode logistic regression ini yaitu memprediksi variabel terkait yang berskala dikotomi. Yang dimaksud dengan skala dikotomi adalah skala nominal yang mempunyai dua kategori, misalkan Ya dan Tidak , atau Tinggi dan Rendah. Dalam persamaan rumus, $\boldsymbol{Pb_j}$ adalah probabilitas yang diprediksi dengan cara dikodekan 1, dan $\boldsymbol{(1\text{-}Pb_j)}$ adalah probabilitas yang diprediksi keputusan lain dengan angka 0 untuk pengkodean [6].

$$\mathbf{Log}\left(\frac{Pb_j}{1-Pb_j}\right)= \alpha + \beta_1.X_{1j} + \beta_2.X_{2j} + \ldots + \beta_n.X_{nj} \qquad (1)$$

Simbol α melambangkan *Intercept,* $X_{1j} \ldots X_{nj}$berarti atribut *independent* dalam catatan *-j,* sedangkan $\beta_1 \ldots \beta_n$ adalah atribut *independent* penurunan, untuk *n* menyatakan jumlah atribut *independent,* sedangkan untuk simbol *j* menyatakan jumlah *record* dalam dataset.

Metode Naive Bayes merupakan salah satu metode machine learning yang menggunakan perhitungan probabilitas. Konsep dasar yang digunakan oleh Bayes adalah Teorema Bayes, yaitu melakukan klasifikasi dengan melakukan perhitungan nilai probabilitas. Klasifikasi dilakukan untuk menentukan kategori dari suatu dokumen. Sebuah keuntungan dari metode Metode Naive Bayes adalah bahwa akan hanya membutuhkan sejumlah kecil data pelatihan untuk memperkirakan parameter (sarana dan varians dari variabel-variabel) yang diperlukan untuk klasifikasi. Karena variabel independen diasumsikan, hanya varians dari variabel untuk masing-masing kelas harus ditentukan dan tidak seluruh matriks kovariansi [11].

Pada umumnya, masalah yang ada di dunia nyata mempunyai bentuk non-linearly separable, sehingga kedua class tidak dapat dipisahkan oleh hyperline secara sempurna. Maka dari itu, diperlukan modifikasi SVM dengan memasukan fungsi kernel. Konsep dari SVM non-linear adlaah mengubah data x yang dipetakan oleh fungsi Φ (x) ke ruang vektor yang memiliki dimensi lebihtinggi. Pemetaan ini bertujuan untuk merepresentasikan dara pada ruang vektor baru [8].

Confusion Matrix adalah sebuah metode yang biasa digunakan untuk perhitungan akurasi. Dalam pengujian keakuratan hasil pencarian akan dievaluasi nilai *recall, precision, accuracy,* dan *error rate*. Dimana *precisi*on mengevaluasi kemampuan sistem untuk menemukan peringkat yang paling relevan, dan didefinisikan sebagai presentase dokumen yang di-*retrieve* dan benar-benar relevan terhadap *query*. *Recall* mengevaluasi kemampuan sistem untuk menemukan semua item yang relevan dari koleksi dokumen dan didefinisikan sebagai presentase dokumen yang relevan terhadap *query*. *Accuracy* merupakan perbandingan kasus yang diidentifikasi benar dengan jumlah seluruh kasus dan *error rate* merupakan kasus yang diidentifikasi salah dengan jumlah seluruh kasus [12].

3. Hasil dan Pembahasan

Pada tahapan ini dilakukan beberapa langkah yaitu dari pengambilan data, pengujian data hingga sampai pada evaluasi atau pembahasan dari hasil yang di dapatkan.

3.1. Pengambilan Data

Pada dataset tersebut terdapat tiga variabel yang akan diujikan, variabel tersebut adalah *"title of review","review text",*dan *"recommendation ID"*. Pada tabel data terdapat kolom *"recommendation ID"* yang berisikan keterangan "1" dan "0" dimana berarti bahwa angka 1 menyatakan bahwa kalimat tersebut

melabelkan *"Recommended"* sedangkan angka 0 menjelaskan bawah kalimat yang terdapat pada baris tersebut adalah "*NotRecommended*". Berikut pada gambar 2 contoh dari data yang akan digunakan untuk penelitian.

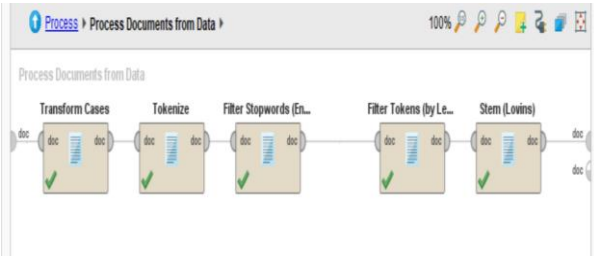
Tabel 1. Dataset *Woman Clothing Ecommerce Review*

No	title of the review	review text	Recommended ID
1	My favorite buy!	I love, love, love this jumpsuit. it's fun, flirty, and fabulous! every time i wear it, i get nothing but great compliments!	Recommended
2	Flattering shirt	This shirt is very flattering to all due to the adjustable front tie. it is the perfect length to wear with leggings and it is sleeveless so it pairs well with any cardigan. love this shirt!!!	Recommended

Data yang digunakan untuk penelitian adalah 4000 data, dengan melalui tahapan pengambilan jumlah komentar secara random dari keseluruhan data. Hal tersebut dilakukan karena keterbatasan *tools* yang digunakan untuk penelitian ini. Selain dengan mengambil sejumlah 4000 data secara acak, hal yang dilakukan terlebih dahulu adalah merapikan data tersebut, dimana terdapat beberapa baris dari variabel yang kosong namun terdapat label yang menyatakan 1 atau 0, maka dilakukan tindakan untuk menghapus dari data yang akan diujikan.

3.2 Persiapan Data (Pre-Processing Data)

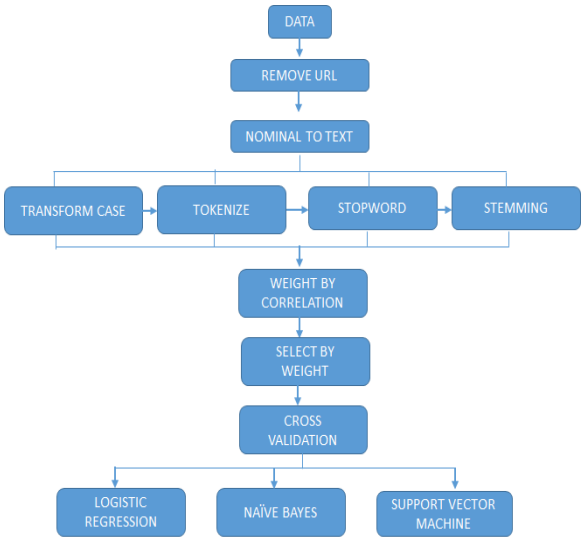
Setelah melalui tahap pembersihan data dan pembagian atribut serta label diatas, proses selanjutnya yang akan dilakukan adalah tahapan *remove url*, dimana setiap kalimat yang mengandung *url* akan dihapus. Tahapan selanjutnya merupakan proses dari *transform case*, *tokenize*, *stopword* hingga *stemming* menggunakan bahasa inggris yang sesuai dengan data yang akan diolah. Proses tersebut dijadikan satu pada *process document from data* yang ada pada rapidminer.



Gambar 2. Process Document From Data

3.3 Pengujian Data

Sehingga dapat dilihat pada gambar 3 untuk alur proses tersebut.



Gambar 3. Alur Proses Pengujian Data

Dalam pengujian data yang pertama dengan algoritma *Logistic Regression* dilakukan dengan menggunakan *weight by correlation*. Seleksi fitur menggunakan *weight by correlation.* Proses tersebut dilakukan sebelum melalui proses *cross validation.* Pada tahapan seleksi fitur dilakukan pada operator *select by weight* dengan *weight relation* dengan top k dimana jumlah k = 100. Dari proses tersebut dijelaskan bahwa pada *weight by correlation* dimana pemilihan kata yang terdapat pada setiap kalimat atau komentar yang ada pada dataset dilakukan pembobotan, kemudian setelah itu diurutkan atau dirangking berdasarkan bobot yang telah ada. Proses pengurutan atau ranking ini diambil dari top k = 100 yang berarti bahwa kata yang digunakan adalah 100 terbaik dari pembobotan kata.

Kemudian dalam menentukan nilai kegunaan dari model yang telah berhasil dibuat pada langkah sebelumnya. Menggunakan *10-fold cross validation* untuk dapat menghasilkan nilai akurasi sebagai bahan perbandingan dalam menentukan algoritma terbaik yang digunakan. *10-fold cross validation* bekerja dengan membagi *dataset* masukan menjadi 10 bagian yang sama rata. 9 bagian kemudian di-training sedangkan yang 1 bagian lainnya digunakan untuk testing. Proses ini diulang sebanyak 10 kali untuk setiap bagian sehingga setiap bagian dari kesepuluh bagian pernah menjadi data untuk testing. Operator *Cross Validation* melakukan proses *10-fold cross validation* ini untuk ketiga algoritma yang digunakan. Untuk setiap percobaan akan dihitung akurasinya. Akurasi akhir adalah nilai rata-rata dari akurasi sepuluh percobaan tersebut. Hasilnya dapat disajikan dalam bentuk *confusion matrix.*

Penggunaan rapidminer secara keseluruhan untuk proses algoritma gambar 3. Sebagaimana dilakukan pula untuk proses pengujian untuk model *Naïve Bayes* dan *Support Machine Vector* dengan proses *cross validation* yang berbeda sesuai dengan model yang digunakan.

3.5. Hasil Penelitian

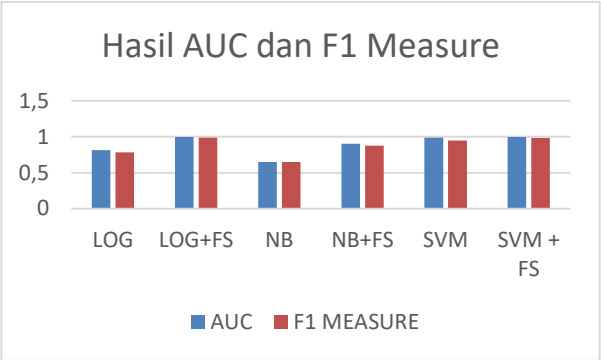
Setelah melalui proses seperti yang diterapkan pada masing-masing algoritma dapat diperoleh hasil akurasi, precision, recall, AUC dan F1 Measure. Nilai Akurasi digunakan sebagai langkah awal perbandingan pencarian algoritma terbaik yang akan dihasilkan. Pebandingan hasil perhitungan nilai akurasi untuk metode *Logistic Regression, Naïve Bayes, dan Support* dapat dilihat pada Tabel 2.

Tabel 2.Hasil Pengujian Data dengan Logistic Regression, Naïve Bayes, dan Support Vector Machine

	LOG	LOG+ FS	NB	NB+F S	SVM	SVM+ FS
Akurasi	**76,72 %**	**98,80 %**	**68,15 %**	**87,97 %**	**94,65 %**	**98,12 %**
Precision	**76,39 %**	**98,71 %**	**71,13 %**	**89,81 %**	**92,67 %**	**97,31 %**
Recall	**78,60 %**	**98,90 %**	**59,70 %**	**85,75 %**	**97,25 %**	**99,00 %**
AUC	**0,816**	**0,999**	**0,651**	**0,905**	**0,989**	**0,999**
F1 Measure	**78,47 %**	**98,80 %**	**64,87 %**	**87,71 %**	**94,90 %**	**98,14 %**

Pada tabel 1 dapat diuraikan dengan hasil perolehan algoritma *Logistic Regression* tanpa seleksi fitur adalah 76,72 persen dan dengan menggunakan seleksi fitur diperoleh 98.80 persen. Dimana hal tersebut mengalami kenaikan pada akurasi. Kemudian untuk hasil perolehan algoritma *Naïve Bayes* tanpa seleksi fitur adalah 68,15 persen dan dengan menggunakan seleksi fitur diperoleh 87,97 persen. Dimana hal tersebut juga mengalami kenaikan pada akurasi setelah adanya penggunaan seleksi fitur. Selanjutnya untuk hasil perolehan algoritma *Support Vector Machine* tanpa seleksi fitur adalah 94,65 persen dan dengan menggunakan seleksi fitur diperoleh 98,12 persen. Hal tersebut juga mengalami kenaikan pada akurasi setelah penggunaan seleksi fitur. Sehingga dari masing-masing algoritma yang telah diujikan mengalami kenaikan akurasi setelah adanya penggunaan seleksi fitur berbasis *pearson correlation* atau pada rapidminer dengan menggunakan *weight by correlation*.

Pada gambar 2 terdapat perbandingan hasil AUC dan F1 Measure. Pada metode *Logistic Regression* tanpa seleksi fitur menghasilkan AUC sebesar 0,816 dan F1 Measure sebesar 78,47% dengan seleksi fitur mengalami kenaikan menjadi 0,999 dan 98,80%. Kemudian untuk metode Naïve Bayes tanpa menggunakan seleksi fitur menghasilkan AUC sebesar 0,651 dan F1 Measure 68,87% setelah dilakuka uji coba dengan menggunakan seleksi fitur mengalami kenaikan menjadi 0,905 dan 87,71%. Pada metode *Support Vector Machine* juga demikian mengalamin kenaikan dari hasil tanpa seleksi fitur AUC sebesar 0,989 dan F1 Measure sebesar 94,90% menjadi 0,999 dan 98,14% setelah adanya penggunaan seleksi fitur.



Gambar 4. Hasil AUC dan F1 Measure

4. Kesimpulan

Akurasi terhadap *opinion mining review pelanggan* mendapatkan hasil akurasi lebih baik (mengalami kenaikan) dengan adanya seleksi fitur berbasis *pearson correlation* terhadap masing masing metode *Logistic Regression, Naïve Bayes, dan Support Vector Machine*. Dari hasil pengujian tersebut dapat dijelaskan bahwa algoritma *Support Vector Machine* adalah algoritma terbaik dari ketiga algoritma yang dibandingkan. Hal ini dapat dilihat pada hasil uji yang dilakukan tanpa menggunakan seleksi fitur. Pada saat uji dengan menggunakan seleksi fitur *weight by correlation* selisih antara kedua algoritma *Support Vector Machine* dan *Logistic Regression* tidak begitu banyak, selisih tersebut mencapai 0,68 % yang didapatkan. Namun ketika diuji tanpa menggunakan seleksi fitur *weight by correlation* hasil algoritma dari *Support Vector Machine* dan *Logistic Regression* memiliki selisih yang lebih banyak yaitu 17,93% dimana algoritma *Support Vector Machine* lebih banyak, dan untuk *Logistic Regression* mengalami kenaikan dengan jumlah yang cukup banyak setelah dilakukan pengujian dengan menggunakan seleksi fitur.

Daftar Rujukan

[1] Widiyanto, I., & Prasilowati, S. L. (2015). Perilaku Pembelian Melalui Internet. *Jurnal Manajemen Dan Kewirausahaan (Journal of Management and Entrepreneurship)*, *17*(2), 109–112. https://doi.org/10.9744/jmk.17.2.109-122.

[2] Agustina, L., & Fayardi, A. O. (2019). *Online Review : Indikator Penilaian Kredibilitas Online dalam Platform E-commerce*. (4), 141–154.

[3] Kusumasondjaja, S., Shanka, T., & Marchegiani, C. (2012). *Journal of Vacation Marketing*. https://doi.org/10.1177/1356766712449365

[4] C, A. R., Lukito, Y., Informatika, P. T., Informasi, F. T., Kristen, U., & Wacana, D. (2017). *Deteksi Komentar Spam Bahasa Indonesia Pada Instagram Menggunakan Naive Bayes*. *IX*(1).

[5] Asghar, M. Z., Kundi, F. M., Khan, A., & Ahmad, S. (2014). Lexicon-Based Sentiment Analysis in the Social Web. *J. Basic. Appl. Sci. Res.*

[6] Zubrinic, K., SJEKAVICA, T., MILICEVIC, M., & OBRADOVIC, I. (2018). A Comparison of Machine Learning Algorithms in Opinion Polarity Classification of Customer Reviews. *International Journal of Computers*, *3*, 159–163.

[7] Wen, H., & Zhao, J. (2017). Aspect term extraction of E-commerce comments based on model ensemble. *2016 13th International Computer Conference on Wavelet Active Media Technology and Information Processing, ICCWAMTIP 2017*, *2018-February*,24–27. https://doi.org/10.1109/ICCWAMTIP.2017.8301421

[8] Purwanto, D. D., & Santoso, J. (2015). *Multinomial Naïve Bayes Classifier Untuk Menentukan Review*. (March), 117–122. Retrieved from https://www.researchgate.net/publication/319256329%0AMULTINOMIAL

[9] Rozy, F., Rangkuti, S., Fauzi, M. A., Sari, Y. A., Dewi, E., & Sari, L. (2018). Analisis Sentimen Opini Film Menggunakan Metode Naïve Bayes dengan Ensemble Feature dan Seleksi Fitur Pearson Correlation Coefficient. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *2*(12), 6354–6361.

[10] Sharma, A., & Dey, S. (2012). Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis. *International Journal of Computer Applications*, (June), 15–20. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Performance+Investigation+of+Feature+Selection+Methods+and+Sentiment+Lexicons+for+Sentiment+Analysis#0

[11] Sugiyono. (2013). Metode Penelitian Pendidikan Pendekatan Kuantitatif, Kualitatif, dan R&D. Bandung: Alfabeta.

[12] Shardlow, M. (2016). An Analysis of Feature Selection Techniques. *The University of Manchester*, (1), 1–7. Retrieved from http://syllabus.cs.manchester.ac.uk/pgt/2018/COMP61011/goodProjects/Shardlow.pdf%0Ahttps://studentnet.cs.manchester.ac.uk/pgt/COMP61011/goodProjects/Shardlow.pdf%0Ahttp://ro.utia.cz http//poseidon.csd.auth.gr%0Ahttp://clopinet.com/isabelle/Projects/NIPS200