Sentiment Analysis Menggunakan Metode Support Vector Machine dan Seleksi Fitur Query Expansion Ranking

Ayu Mashita Hardiyanti [1], Yuliant Sibaroni [2]
[1,2] Prodi S1 Informatika, Fakultas Informatika, Universitas Telkom, Bandung
[1]ayumashita@students.telkomuniversity.ac.id,
[2]yuliant@telkomuniversity.ac.id,

Abstrak

Pertumbuhan pengguna aktif di media sosial sangat berkembang pesat. Pengguna aktif media sosial sering mengutarakan pendapatnya terhadap sebuah layanan atau produk melalui media sosial ternama seperti tweeter, Instagram, Tripadvisor, sehingga pendapat atau ulasan sangat banyak ditemukan pada media sosial. Ulasan dapat dijadkan sebagai penilaian penting dan bermanfaat jika dikelola dengan baik. Membaca ulasan yang banyak di media sosial membutuhkan waktu yang cukup lama, maka dari itu membutuhkan klasifikasi sentiment yang dapat mengelompokkan menjadi dua kelas yaitu kelas positif dan negative. Metode klasifikasi yang digunakan adalah Support Vector Machine yang memiliki kemampuan untuk menerapkan pemisah linear pada input data non linear berdimensi tinggi yang diperoleh dengan menggunakan fungsi kernel yang dibutuhkan. Untuk mendukung suatu penelitian agar lebih maksimal, terdapat seleksi fitur yang akan digunakan untuk mereduksi fitur-fitur sehingga proses kalsifikasi lebih efektif dan efisien. Seleksi fitur yang akan digunakan adalah Query Expansion Ranking yang dapat memaksimalkan hasil akurasi. Hasil yang di dapatkan dari penilitian ini adalah nilai akurasi tertinggi dengan penggunaan kernel Polinomial dan RBF pada penggunaan Rasio 75%.

Kata kunci: support vector machine, query expansion ranking, sentiment analysis, seleksi fitur, klasifikasi teks

Abstract

The growth of active users on social media is growing rapidly. Active social media users often express their opinion on a service or product through well-known social media such as tweeters, Instagram, Tripadvisor, so that opinions or reviews are very much found on social media. Reviews can be an important and useful assessment if managed properly. Reading a lot of reviews on social media takes a long time, therefore it requires a sentiment classification which can be grouped into two classes, namely positive and negative classes. The classification method used is the Support Vector Machine which has the ability to apply a linear separator to high-dimensional non-linear data input obtained by using the required kernel functions. To support a research to be more optimal, there is a selection of features that will be used to reduce features so that the calcification process is more effective and efficient. The feature selection that will be used is Query Expansion Ranking which can maximize accuracy results. The results obtained from this study are the highest accuracy values with the use of the Polynomial kernel and RBF kernel use Rasio 75% .

Keywords: support vector machine, query expansion ranking, sentiment analysis, feature selection, text classification

1. Pendahuluan

Pertumbuhan pengguna aktif di media sosial sangat berkembang pesat. Dari tahun ke tahun jumlah pengguna internet semakin banyak. Berdasarkan laporan DATAREPORTAL pengguna internet di Indonesia pada bulan Januari 2021 mencapai 202,6 juta dan meningkat lebih dari 16% dari tahun 2020 [1]. Perkembangan teknologi ini memiliki dampak postif seperti memudahkan dalam hal komunikasi, mencari dan mengakses sebuah informasi. Pengguna aktif internet sering mengutarakan pendapat mereka terhadap sebuah layanan atau produk melalui media sosial ternama seperti Twitter, Facebook, Instagram, Tripadvisor, dan lain sebagainya, sehingga pendapat atau ulasan sangat banyak ditemukan di internet.

Ulasan dapat dijadikan sebagai penilaian yang penting untuk beberapa faktor jika dikelola dengan baik dan bermanfaat bagi bidang usaha dan upaya pemasaran. Dengan adanya ulasan ini, dapat digunakan sebagai referensi dan pengambilan keputusan. Pengguna internet bergantung pada sebuah rekomendasi word-of-mouth atau opini sebelum menggunakan sebuah produk, dikarenakan pentingnya sebuah review dari pengguna lain bisa memberikan informasi tentang produk tersebut berdasarkan perspektif pengguna lain yang sudah pernah menggunakan produk tersebut [2].

Ulasan mengenai suatu penilaian dapat berupa ulasan positif dan negative. Dengan membaca semua ulasan untuk mendapatkan sebuah informasi membutuhkan waktu yang cukup lama, maka dari itu dibutuhkan sebuah metode yang bisa digunakan untuk mendapatkan sebuah informasi pada ulasan secara efektif dan efisien, salah satunya menggunakan sebuah metode Text Mining. Analisis sentiment merupakan

salah satu teknik pada Text Mining. Analisis sentiment adalah studi komputasi mengenai sikap, emosi, pendapat, penilaian, pandangan dari sekumpulan teks [3]. Dengan adanya analisis sentiment informasi yang didapatkan bisa menjadi landasan sebuah perusahaan untuk melakukan sebuah inovasi atau perbaikan.

Berdasarkan uraian diatas, perlu dilakukan anilisis lanjutan untuk mengetahui sebuah opini. Opini atau Ulasan akan diklelompokkan pada ulasan positif dan negative, dengan menggunakan pendekatan klasifikasi pada sentiment analisis. Metode yang akan digunakan yaitu metode Support Vector Machine (SVM). Metode SVM dipilih dikarenakan mampu menemukan hyperline terbaik sebagai pemisah [4]. SVM memiliki akurasi yang baik dalam klasifikasi, dan proses learning yang cepat [5], metode SVM memiliki nilai Area Under Curve (AUC) terbesar dibandingkan metode Neural Network, KNN, dan Decision Tree saat mengklasifikasikan ulasan pariwisata Indonesia pada media sosial Twitter [6].

Untuk mendukung suatu penelitian agar lebih maksimal, terdapat seleksi fitur yang akan digunakan untuk mereduksi fitur-fitur sehingga proses kalsifikasi lebih efektif dan efisien. Penelitian mengenai seleksi fitur pernah diterapkan pada penelitian Parlar dkk [7] yang menerapkan seleksi fitur *Query Expansion Ranking* (QER) untuk analasis sentiment empat kategori, dan membanding dengan seleksi fitur lain seperti *Chi_Square* dan *Document Frequency Difference*. Hasil dari penelitian menunjukkan bahwa seleksi fitur Query Expansion Ranking memiliki nilai akurasi yang baik dibanding dengan seleksi fitur *Chi_Squere* dan *DFD*. Penelitian lain tentang Seleksi Fitur QER juga diterapkan pada penelitian sentiment analisis pariwisata kota Malang, penelitian tersebut memiliki tujuan mengetahui opini mengenai pariwisata kota Malang. Penelitian ini menggunakan metode Naïve Bayes Classifier dengan *feature selection* QER [8]. Berdasarkan penelitian sebelumnya dapat diketahui bahwa penggunaan Seleksi Fitur Quer Expansion Ranking merupakan seleksi fitur yang tepat untuk digunakan pada penelitian ini.

Berdasarkan uraian dari penelitian sebelumnya, penggunaan algoritma klasifikasi Naïve Bayes belum mendapatkan hasil akurasi yang maksimal dalam menyelesaikan permasalahan yang dilakukan, sehingga akan dilakukan penelitian yang dapat menutupi kekurangan dari metode Naïve Bayes Classifier dan meningkatkan akurasi dari penggunaan metode SVM dan seleksi fitur QER. Kemudian penulis akan membuktikan apakah metode Query Expansion Ranking dan Support Vector Machine dapat diimplementasikan dan mendapatkan hasil akurasi yang baik, dan apakah penggunaan model kernel pada Support Vector Machine memiliki pengaruh terhadap hasil kalsifikasi.

Tujuan Penelitian ini yang pertama adalah untuk mengetahui efektifitas Query Expansion Ranking terhadap metode klasifikasi *Support Vector Machine*. Yang kedua untuk mengetahui perbandingan model kernel pada Support Vector Machine.

Batasan masalah pada penelitian ini yaitu data tempat wisata di Pulau Madura dengan menggunakan ulasan berbahasa Indonesia. Metode penelitian ini hanya menggunakan metode *Support Vector Machine* dan seleksi fitur *Query Expansion Ranking*.

2. Studi Terkait

2.1 Penelitian Terkait

Analisis sentiment terhadap tokoh publik dengan menggunakan algoritma support vector machine. Support vector machine merupakan sebuah teknik machine learning yang sering digunakan untuk klasifikasi teks. Data yang digunakan pada penelitian ini diambil dari twitter yang dipisahkan menjadi data training dan data testing. Data yang telah didapatkan akan diolah dengan melakukan text preprocessing kemudian diklasifikasikan dengan menggunakan algoritma support vector machine. Pada algoritma SVM ini terdapat 4 kernel yaitu kernel linear, kernel polynomial, kernel sigmoid, dan kernel Gaussian. Kinerja algoritma ini dievaluasi dengan menggunakan nilai presisi, akurasi, dan recall. Hasil yang didapatkan pada penelitian ini adalah menunjukkan bahwa keral linear memiliki nilai presisi yang baik dengan hasil nilai 80%, kernel sigmoid memiliki nilai recall dan nilai akurasi yang paling baik dengan hasil nilai recall 85% dan hasil nilai akurasi 81% [9].

Pada penelitian analisa sentiment review hotel menggunakan algoritma Support Vector Machine berbasis Particle Swarm Optimization. Analisa sentiment atau opinion mining merupakan salah satu solusi mengatasi masalah untuk mwngwlompokan opini atau review secara otomatis. Metode yang digunakan pada penelitian ini adalah Support Vector Machine dan seleksi fitur Particle Swarm Optimization (PSO). Particle Swarm Optimization merupakan suatu teknik optimasi yang sangat sederhana untuk menerapkan dan memodifikasi beberapa parameter. Pada pengujian ini nilai akurasi yang didapatkan akan menjadi tolak ukur untuk mencari model pengujian terbaik. Hasil dari penelitian ini menunjukkan nilai akurasi sebesar 96,94% [10].

Pada penelitian Analisis Sentiment terhadap wacana politik pada media masa online menggunakan algortima SVM dan Naïve Bayes. Naïve Bayes merupakan salah satu metode klasifikasi yang memiliki kelebihan yaitu algoritma yang sederhana dan memiliki kecepatan yang tinggi dalam proses pelatihan dan klasifikasi. Pada penelitian ini akan membandingkan metode Naïve Bayes dengan Support Vector Machine. Hasil eksperimen menunjukkan bahwa menguji data sentimen analisis wacana politik media masa online mengunakan algoritma Naive Bayes hasilnya mendapatkan akurasi sebesar 59,98 % dan membandingkan dengan algoritma Support Vector Machine dengan hasil yang lebih baik yaitu 90,50% [11].

Pada penelitian sentiment analysis pada teks bahasa indonesia menggunakan SVM dan KNN. K-Nearest Neighbor (KNN) memiliki tujuan mengklasifikasikan obyek berdasarkan atribut dan training sample. Peneliian ini menggunakan teks bahasa indonesia kemudian metode KNN akan mengklasifikasi langsung pada data pembelajaran agar dapat menentukan model yang akan dibentuk oleh SVM untuk menentukan kategori dari data baru yang ingin ditentukan kategorinya. Hasil yang didapatkan pada penelitian ini adalah nilai akurasi yang tertinggi didapatkan saat menggunakan metode SVM yaitu dengan nilai 67,90% dan nilai akurasi saat menggunakan metode KNN yaitu 60,20% [12].

Pada penelitian analisis sentimen kurikulum 2013 pada sosial media menggunakan metode KNN dan seleksi fitur Query Expansion Ranking. Seleksi Fitur dibutukan untuk membantu mempercepat proses komputasi, pengklasifikasian menjadi lebih efisien selain itu dapat membantu mengoptimalkan nilai akurasi klasifikasi karena menghilangkan fitur noise[13]. Pada penelitian ini memiliki dua skenario yaitu yang pertama adalah pengujian variasi nilai K pada klasifikasi KNN dan mendaptkan hasil K yang terbaik adalah pada saat K = 1. Kemudian pengujian kedua adalah pengujian Rasio pada Seleksi Fitur yang menggunakan perbandingan Rasio 25%, 50%, 75%, dan 100% dan mendapatkan rasio terbaik pada saat rasio 50% dengan hasil akurasi sebesar 96,36% [14].

Pada penelitian Analisis Sentiment pariwisata kota Malang menggunakan metode Naïve Bayes dan seleksi fitur Query Expansion Ranking. Proses dari penelitian ini terdiri dari preprocessing, seleksi fitur QER, dan klasifikasi dengan Naïve Bayes. Seleksi fitur QER ini berfungsi untuk mengurangi jumlah fitur pada saat proses klasifikasi. Pengujian pada penelitian ini adalah uji akurasi dengan menggunakan variasi rasio seleksi fitur QER dengan mengganti rasio antara 0 – 100. Berdasarkan pengujian algoritma Naïve Bayes dan QER menghasilkan akurasi tertinggi 86,6% pada penggunaan rasio 75% [8].

Table 1 Penelitian Terdahulu

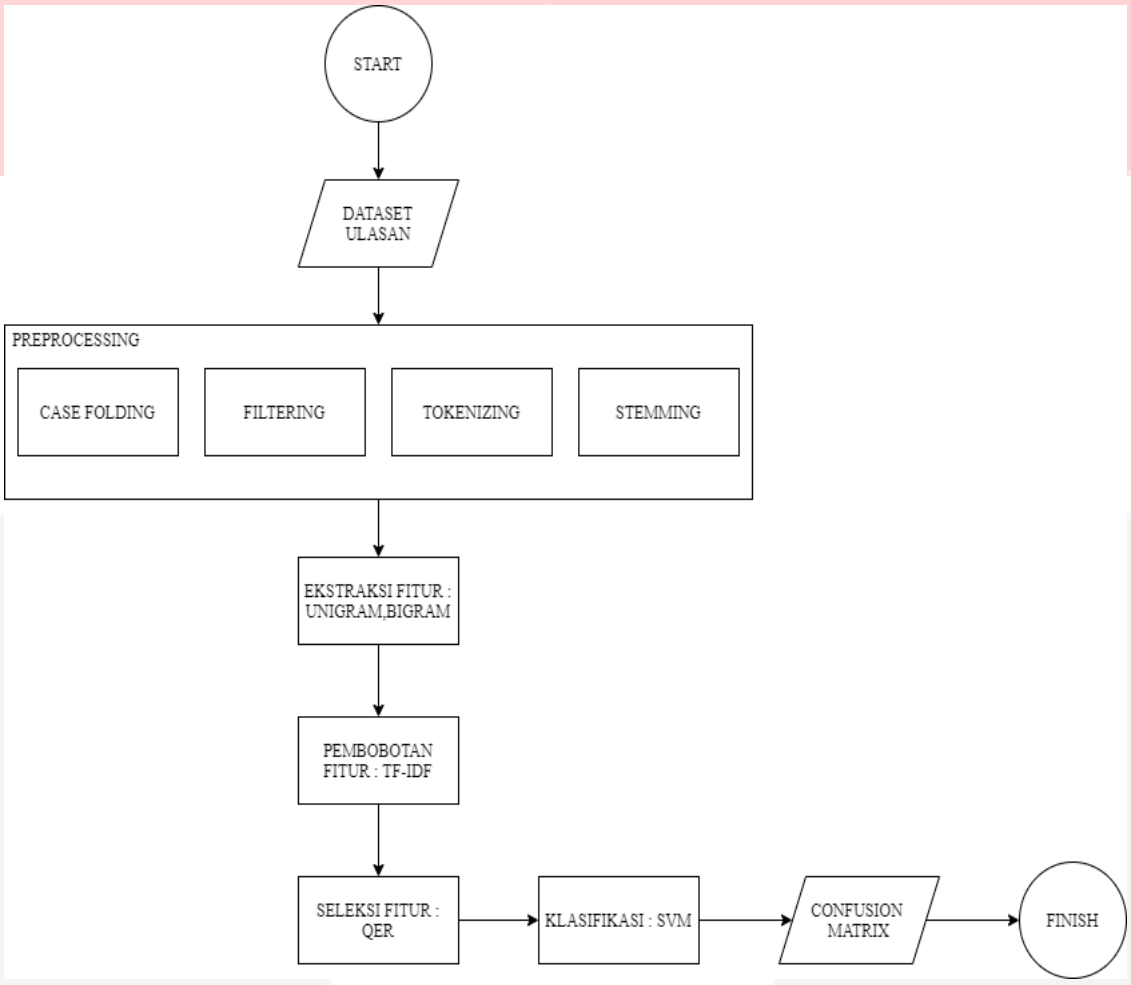
PENULIS	METODE KLASIFIKASI	METODE SELEKSI FITUR	HASIL	KETERKAITAN
Taufik dan S.A Pamungkas (2018)	*Support Vector Machine*	-	Kernel linear = 77% Kernel polinom = 50% Kernel Gaussian = 50% Kernel Sigmoid = 81%	Pada penelitian yang dilakukan penulis akan menggunakan metode *Support Vector Machine* dengan seleksi fitur Query Expansion Ranking yang akan membandingkan 4 parameter kernel pada SVM.
Elly Indrayuni (2016)	*Support Vector Machine*	Particle Swarm Optimizati	96,94%	Pada penelitian yang dilakukan penulis akan menggunakan metode

		on		*Support Vector Machine* dengan seleksi fitur Query Expansion Ranking
Andi Nurul Hidayat (2015)	*Support Vector Machine* dan Naïve Bayes	-	SVM = 90,50% Naïve Bayes = 59,98%	Pada penelitian yang dilakukan penulis akan menggunakan metode SVM dikarenakan metode SVM merupakan salah satu algoritma yang paling akurat dengan menghasilkan nilai akurasi sebesar 90,50%.
Syahfitri Kartika Lidya, Opim Salim Sitompul, Syahril Efendi (2015)	*Support Vector Machine* dan K-Nearest Neighbor	-	SVM = 67,90% KNN = 60,30%	Pada penelitian yang dilakukan penulis akan menggunakan metode SVM dikarenakan metode SVM merupakan salah satu algoritma yang paling akurat dengan menghasilkan nilai akurasi sebesar 67,90%
Nurul Dyah Mentari, M. Ali Fauzi, Lailil Muflikhah (2018)	K-Nearest Neighbor	Query Expansion Ranking	96,36%	Pada penelitian yang dilakukan oleh penulis akan menggunakan seleksi fitur QER dikarenakan penggunaan seleksi fitur dapat membantu meningkatkan hasil akurasi, terlihat pada penelitian yang dilakukan oleh Nurul dkk menghasilkan nilai akurasi 96,36%.
Shima Fannisa, M. Ali Fauzi, Sigit Adinugroho (2018)	Naïve Bayes	Query Expansion Ranking	86,6%	Pada penelitian yang dilakukan oleh penulis akan menggunakan seleksi fitur QER dikarenakan penggunaan seleksi fitur dapat membantu meningkatkan hasil akurasi, terlihat pada penelitian yang dilakukan oleh Shima dkk menghasilkan nilai akurasi 86,6%.

Berdasarkan Tabel 1 penulis menggunakan penelitian sebelumnya sebagai refrensi untuk membantu dalam melakukan riset yang sedang dilakukan. Pada setiap hasil penelitian sebelumnya, menunjukkan bahwa riset yang dilakukan pada penelitian ini melakukan klasifikasi ulasan menggunakan seleksi fitur Query Expansion Ranking dan metode klasifikasi *Support Vector Machine* yang belum digunakan pada penelitian-penelitian sebelumnya.

3. Sistem yang Dibangun

System yang akan buat pada penelitian ini menggunakan metode *Support Vector Machine* dan seleksi fitur Query Expansion Ranking merupakan system yang mampu melakukan serangkaian proses yang dapat mengklasifikasi sebuah ulasan menjadi dua kelas yaitu kelas positif dan negatif. System ini melakukan beberapa tahap yang digambarkan pada Gambar 1 berikut.



Gambar 1 Proses Klasifikasi

3.1 Dataset Ulasan

Data yang digunakan pada penelitian ini adalah data ulasan Pariwisata Pulau Madura yang diperoleh melalui tahap scrapping dari situs web TripAdvisor. Pelabelan data dilakukan secara manual dengan menggunakan bantuan rating pada setiap ulasan, ulasan dengan rating kurang dari 3 maka ulasan tersebut masuk pada kelas negative (-1), ulasan dengan rating samadengan 3 maka ulasan tersebut masuk pada kelas netral (0), dan ulasan dengan rating diatas 3 maka ulasan tersebut masuk pada kelas positif.

Pada penelitian ini hanya terdapat dua kelas yang digunakan, yaitu kelas positif (1) dan kelas negative (-1). Pada ulasan kelas positif mengandung kalimat atau pernyataan yang positif seperti kalimat pujian, terimakasih, sanjungan, dll. Ulasan kelas negative memiliki kalimat atau pernyataan yang negative seperti cacian, penghinaan, dll.

Ulasan kelas netral akan dilakukan reduksi kelas dengan mengkategorikan kelas sentiment netral positif dan negative dengan cara manual. Ketentuan yang digunakan adalah apabila kelas sentiment netral tidak teridentifikasi kata negative maka ulasan tersebut akan masuk pada kelas positif, dan apabila kelas netral memiliki kata negative maka ulasan tersebut akan masuk pada kelas negative. Pada Tabel 2 merupakan contoh dataset yang akan digunakan pada penelitian ini.

Table 2 Contoh Dataset

No	Ulasan	Rating	Label
1	suasana pantai agak kotor, mungkin karena musim lebaran.	1	-1
2	Museum itu sangat kecil mungkin hanya berukuran 90 meter az. Baru masuk, berjalan sebentar sudah habis semua apa yang mau qta lihat.	2	-1

3	Surga di Sumenep itu adalah Pulau Gili Labak dengan balutan pasir putih dan bawah airnya yang begitu cantik membuat tak ingin segera pulang.	5	1
4	Ini adalah benar-benar tempat yang indah. Tenang Incredibely. Kami mengunjungi pulau ini selama akhir pekan, jadi itu sedikit sibuk di port, tapi tak seorang pun tampaknya tertarik untuk berjalan-jalan di sekitar pulau. Keluarga kami tinggal di sana hanya 2 jam karena kelompok kami. Mudah-mudahan kami akan dapat kembali lagi.	5	1
5	Tempat wisata yang cocok dikunjungi bersama keluarga dan anak-anak, banyak benda bdenda bersejarah peninggaalan kerajaan Sumenep masih tersimpan di sini	4	1
6	bila kesini jangan lupa bawa syal, masker topi dan kacamata hitam karena sangat panas saat siang hari dan anda berasa seperti berada di gurun seperti yang ada di negara timur tengah	3	1
7	Bagaimanapun, ketika saya mengunjungi danau, tidak ada air sama sekali. Mereka mengatakan itu tergantung pada hujan. betapa mengecewakan, sangat sulit untuk pergi ke bukit Jaddih dan pada saat kami tiba, tidak ada air sama sekali. Foto-foto yang terlihat di sini dan di Google (dengan air hijau) hanya akan terjadi setelah hujan monsun. Karena perubahan iklim, hujan belum turun selama berbulan-bulan, dan kering. Anda dapat melihat perahu yang ditinggalkan dan benar-benar tidak ada yang lain.	3	-1

Setelah melakukan pelabelan pada setiap ulasan maka data akan dihitung total data sebanyak 496 data dengan perbandingan kelas negative sebanyak 54 data dan positf sebanyak 442 data. Data yang didapatkan merupakan data imbalanced. Data Imbalanced merupakan data yang memiliki jumlah data yang tidak seimbang antara satu kelas dengan kelas lainnya. Data mining mengartikan imbalanced merupakan kondisi ketika jumlah data pada kelas mayoritas lebih banyak dibandingkan dengan jumlah data pada kelas minoritas [15]. Penggunaan data imbalanced pada klasifikasi akan menimbulkan masalah ketika perbandingan kelas positif dan negative sangat besar, sehingga akan mempengaruhi hasil akurasinya. Sehingga pada penelitian ini menggunakan metode resampling untuk mengatasi masalah imbalanced. Hasil yang didapatkan setelah melakukan resampling adalah kelas negative sebanyak 442 data dan posotif sebanyak 442 data dengan total keseluruhan data adalah 884 data.

Selanjutnya data yang telah melalui tahap resampling akan dibagi menjadi data training dan data testing. Data training digunakan untuk membentuk model klasifikasi dan merupakan representasi knowledge yang akan digunakan untuk memprediksi kelas data baru. Data testing digunakan untuk mengukur performa dari model yang didapatkan. Semakin banyak data training yang digunakan maka akan semakin banyak belajar sehingga ketelitian akan semakin baik [16]. Sehingga pada penelitian ini menggunakan perbandingan rasio data training dan data testing sebesar 80:20.

3.2 Preprocessing

Tahap prepocessing data dapat juga disebut sebagai Ekstraksi Data merupakan sebuah implementasi text mining yang digunakan untuk memproses sebuah teks menjadi lebih terstruktur. Prepocessing dilakukan dengan cara mengeliminasi data yang tidak sesuai agar mudah di proses oleh sistem. Pada penelitian ini tahap Prepocessing yang digunakan yaitu:

a. Case Folding

Proses Case Folding yaitu menyeragamkan bentuk huruf –huruf mulai “a” sampai “z”, huruf-huruf kapital diubah menjadi huruf kecil. Berikut ilustrasi proses Case Folding yang ditunjukkan pada tabel dibawah ini:

Table 3 Contoh Proses Case Folding

Sebelum Case Folding	Sesudah Case Folding
Pantainya yang sangat indah dan nyaman, kita bisa menikmati snorkling sepuasnya dengan mengelilingi pulau.	pantainya yang sangat indah dan nyaman, kita bisa menikmati snorkling sepuasnya dengan mengelilingi pulau.

b. Filtering

Proses Filtering yaitu pengecekan pada kamus untuk menghilangkan kata-kata seperti kata konjungsi sesuai dengan kamus stopword removal dan menghilangkan karakter yang tidak diperlukan. Berikut ilustrasi

proses Filtering yang ditunjukkan pada tabel dibawah ini:

Table 4 Contoh Proses Filtering

Sebelum Filtering	Sesudah Filtering
pantainya **yang sangat** indah **dan** nyaman kita **bisa** menikmati snorkling sepuasnya **dengan** mengelilingi pulau	pantainya indah nyaman kita menikmati snorkeling sepuasnya mengelilingi pulau

c. Tokenizing

Proses Tokenizing yaitu proses memecahkan kalimat menjadi term-term berdasarkan spasi. Hasil dari proses ini adalah beberapa kumpulan kata-kata tanpa tanda baca, karakter, dan angka. Berikut ilustrasi proses Case Folding yang ditunjukkan pada tabel dibawah ini:

Table 5 Contoh Proses Tokenizing

Sebelum Tokenizing	Sesudah Tokenizing
pantainya indah nyaman kita menikmati snorkeling sepuasnya mengelilingi pulau	“pantainya” “indah” “kita” “menikmati” “snorkling” “sepuasnya” “mengelilingi” “pulau”

d. Stemming

Proses stemming yaitu proses untuk mengganti kata dasar dari sebuah kata dengan cara menghilangkan semua imbuhan (affixes) baik dari awalan (prefixes), sisipan (infixes), akhiran (suffixes) dan kombinasi awalan dan akhiran (confixes) [17]. Pada penelitian ini menggunakan stemming arifin-setiono yang sudah banyak digunakan untuk proses stemming pada teks bahasa indonesia.Berikut ilustrasi proses stemmning yang ditunjukkan pada tabel dibawah ini:

Table 6 Contoh Proses Stemming

Sebelum Stemming	Sesudah Stemming
“pantainya” “indah” “kita” “menikmati” “snorkling” “sepuasnya” “mengelilingi” “pulau”	pantai indah nyaman kita nikmat snorkeling puas keliling pulau

Tahap-tahap prepocessing diatas menggunakan contoh data ulasan yang diperoleh dari website TripAdvisor yang disimpan dalam database. Hasil preprocessing data ulasan dapat dilihat pada tabel 7.

Table 7 Hasil Prepocessing

Sebelum Preprocessing	Setelah preprocessing
Pantainya yang sangat indah dan nyaman, kita bisa menikmati snorkling sepuasnya dengan mengelilingi pulau.	pantai indah nyaman kita nikmat snorkeling puas keliling pulau

3.3 Ekstraksi Fitur : Unigram dan Bigram

Metode N-gram digunakan untuk memprediksi item yang berurutan pada sebuah teks. N-gram merupakan sebuah metode pemotongan yang memecah teks panjang menjadi sebuah teks yang sederhana. N pada N-gram menentukan berapa banyak elemen yang akan dihasilkan dan memecah sebuah kalimat menjadi kecil-kecil sesuai dengan pemberian bobot pada N. Terdapat beberapa jenis N-gram yaitu Unigram, Bigram, dan Trigram [18].

N-gram memiliki keunggulan yaitu untuk menangani input yang tidak jelas dan bisa digunakan untuk hal-hal seperti memulihkan teks. Memecah teks menjadi kecil-kecil juga merupakan slaha satu kelebihan metode N-gram, dikarenakan apabila terdapat kesalahan dalam proses maka hanya sebagian kecil yang terpengaruh sedangkan yang akan tetap utuh [19].

Proses penerapan N-gram yaitu pertama-tama setiap kata pada kalimat akan dipecah menjadi sebuah token pada setiap kata, kemudian token yang berdekatan dikelompokkan sesuai jumlah N pada N-gram. Nilai N yang digunakan pada penelitian ini menggunakan Unigram dan Bigram. Berikut contoh penggunaan proses N-gram.

Table 8 Contoh Ekstraksi Fitur dengan Unigram dan Bigram

Kalimat : pulau yang masih alami pasir putih dan terumbu karang yang exotic
Unigram : “Pulau” “yang” “masih” “alami” “pasir” “putih” “dan” “terumbu” “karang” “yang” “exotic”

Bigram :
"_ pulau" "pulau yang" "yang masih" "masih alami" "alami pasir" "pasir putih" "putih dan" "dan terumbu" "terumbu karang" "karang yang" "yang exotic"

3.4 Pembobotan Fitur : TF-IDF

Term Frequency (TF) merupakan pengukuran sederhana pada metode pembobotan. Pada metode ini masing-masing term diasumsikan mempunya proposisi kepentingan sesuai jumlah munculnya sebuah teks pada dokumen. Ketika term yang frequent cenderung muncul sehingga term-term tersebut memiliki kekuatan pembeda (keunikan) yang kecil maka, TF dapat memperbaiki nilai recall pada information retrieval, tetapi tidak selalu memperbaiki nilai precision. Secara sistematika term frequency yang muncul dalam sebuah dokumen dapat dirumuskan seperti persamaan dibawah ini.

$$TF(t,d) = \sum_{x \in d} f_t(x) \quad (1)$$

Dimana nilai : $f_t(x)$ bernilai 1 jika x = t dan bernilai 0 jika x ≠ t

TF(t,d) merupakan frekuensi dari term t pada teks d

Invers document frequency (IDF) adalah metode pembobotan term yang lebih fokus untuk memperhatikan kemunculan term pada keseluruhan kumpulan teks. Pada IDF term yang jarang muncul pada keseluruhan koleksi teks dinilai lebih berharga. Nilai kepentingan tiap term diamsumsikan berbanding terbalik dengan jumlah teks yang mengandung term tersebut. Sehingga dapat dirumuskan dengan persamaan dibawah ini.

$$IDF_t = \log(\frac{N}{DF_t}) \quad (2)$$

Dimana nilai : DF_t adalah banyaknya dokumen yang membuat t dan N merupakan jumlah total dokumen.

Term Frequency Invers Document Frequency (TFIDF) adalah metode pembobotan yang menggabungkan metode TF dan IDF. Metode ini diusulkan oleh slaton sebagai sebuah kombinasi metode yang dapat memberikan performansi yang lebih baik, khususnya dalam memperbaiki nilai recall dan precision.

Metode pembobotan yang diintregrasikan dari *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) yang dijabarkan dengan rumus :

$$W(t,d) = TF(t,d)*IDF_t \quad (3)$$

Dimana nilai TF(t,d) mewakili *Term Frequency* dan IDF_t mewakili Invers Document Frequency.

Metode ini fungsinya untuk mencari representasi nilai dari kumpulan data yang *training* yang hasilnya dibentuk oleh *vector* antara dokumen dan kata dan dicluster berdasarkan kesamaan antara ulasan.

Berikut ini terdapat 4 sample data yang diterapkan pada rancangan TF-IDF pada penelitian ini:

D1 : pantai indah kunjung tidak nyesal (Ulasan positif)

D2 : sampah serak ganggu nyesal kunjung pantai (Ulasan negative)

D3 : pantai indah tidak polusi pulau kunjung kembali (Ulasan positif)

D4 : bersih pantai ganggu lihat sampah tebar (Ulasan negative)

Table 9 Contoh Perhitungan TF-IDF

Q	tf				DF	IDF+1	W = tf * (IDF +1)			
	D1	D2	D3	D4			D1	D2	D3	D4
Pantai	1	1	1	1	4	Log(4/4) = 0 + 1 = 1	1	1	1	1
Indah	1	0	0	0	1	Log(4/1) = 0,6 + 1 =1,6	1,6	0	0	0
Kunjung	1	1	1	0	3	Log(4/3) = 0,1 + 1 =1,1	1,1	1,1	1,1	0
Tidak	1	0	1	0	2	Log(4/2) = 0,3 + 1 =1,3	1,3	0	1,3	0
Nyesal	1	1	0	0	2	Log(4/2) = 0,3 + 1 =1,3	1,3	1,3	0	0
Sampah	0	1	0	1	2	Log(4/2) = 0,3 + 1 =1,3	0	1,3	0	1,3
Serak	0	1	0	0	1	Log(4/1) = 0,6 + 1 =1,6	0	1,6	0	0
Ganggu	0	1	0	1	2	Log(4/2) = 0,3 + 1 =1,3	0	1,3	0	1,3
Indah	0	0	1	0	1	Log(4/1) = 0,6 + 1 =1,6	0	0	1,6	0
Polusi	0	0	1	0	1	Log(4/1) = 0,6 + 1 =1,6	0	0	1,6	0
Pulau	0	0	1	0	1	Log(4/1) = 0,6 + 1 =1,6	0	0	1,6	0
Kembali	0	0	1	0	1	Log(4/1) = 0,6 + 1 =1,6	0	0	1,6	0
Bersih	0	0	0	1	1	Log(4/1) = 0,6 + 1 =1,6	0	0	0	1,6
Lihat	0	0	0	1	1	Log(4/1) = 0,6 + 1 =1,6	0	0	0	1,6
tebar	0	0	0	1	1	Log(4/1) = 0,6 + 1 =1,6	0	0	0	1,6

3.5 Seleksi Fitur : *Query Expansion Ranking*

Seleksi fitur Query Expansion Ranking digunakan untuk mengurangi fitur yang kurang relevan dan mengoptimalkan kinerja klasifikasi. Query Exapansion Ranking terinspirasi dari metode Query Expansion dan Probabilistic Weighting Model. Query Expansion berguna untuk meningkatkan kualitas query yang diinputkan , kemudian Probabilistic Weighting Model berguna untuk memberikan skor untuk setiap fitur [7]. Berikut persamaan yang menunjukkan proses perhitungan untuk seleksi fitur Query Expansion Ranking:

$$pf = \frac{df_+^f + 0.5}{n^+ + 1.0} \quad (4)$$

Keterangan :

pf = nilai probabilitas term f pada dokumen data bernilai positif

df_+^f = jumlah dokumen yang mengandung term f pada data ulasan yang memiliki nilai positif

n^+ = jumlah seluruh dokumen data bernilai positif

$$qf = \frac{df_-^f + 0.5}{n^- + 0.5} \quad (5)$$

Keterangan :

qf = nilai probabilitas term f pada dokumen data bernilai negatif

Df_-^f = jumlah dokumen yang mengandung term f pada data ulasan yang memiliki nilai negatif

n^- = jumlah seluruh dokumen data bernilai negatif

$$score\ f = \frac{|\ pf + qf\ |}{|\ pf - qf\ |} \quad (6)$$

Keterangan :

score f = hasil perhitungan QER untuk term f.

Penjelasan mengenai alur proses seleksi fitur menggunakan Query Expansion Ranking yaitu :

1. Data yang sudah melalui proses Resampling, Preprocessing, Ekstraksi Fitur Unigram dan Bigram, dan proses Pembobotan TF-IDF,
2. Pada dokumen kelas positif (pf) akan dihitung peluang fitur (f) dengan persamaan (4),
3. Selanjutnya, pada dokumen kelas negative (qf) akan dihitung peluang fitur (f) dengan persamaan (5),
4. Setelah mendapatkan nilai (pf) dan (qf), maka akan dihitung nilai score dari fitur (f) dengan persamaan (6),
5. Hasil dari perhitungan score f akan digukanan untuk proses selanjutnya.

Hasil dari seleksi fitur Query Expansion Ranking akan diurutkan dimulai dari data yang memiliki skor paling besar hingga skor paling kecil. Selanjutnya data yang memiliki skor yang kecil akan dihapuskan dan data dapat digunakan pada proses klasifikasi menggunakan Support Vector Machine. Adapun penjelasan mengenai alur proses tahapan seleksi fitur QER akan di contohkan dengan perhitungan manual sebagai berikut:

Table 10 Data contoh perhitungan manual QER

No	Kata Positif	Jumlah Kata	Kata Negatif	Jumlah kata
1.	pantai	2	pantai	2
2.	kunjung	2	sampah	2
3.	tidak	2	ganggu	2
4.	kembali	1	serak	1
5.	nyesal	1	bersih	1
6.	indah	1	lihat	1
7.	polusi	1	tebar	1
8.	pulau	1	kunjung	1
9.			nyesal	1
	Total Dokumen	**11**		**12**

Setelah jumlah kata pada kelas positif dan negative diketahui maka akan dilakukan perhitungan pada persamaan QER sebagai berikut :

$$pkunjung = \frac{df_+^f + 0.5}{n^+ + 1.0}$$

$$pkunjung = \frac{2 + 0.5}{11 + 1.0} = 0{,}208$$

$$qkunjung = \frac{df_-^f + 0.5}{n^- + 0.5}$$

$$qkunjung = \frac{1 + 0.5}{12 + 0.5} = 0{,}12$$

Diketahui nilai pkunjung = 0,208 dan nilai qkunjung = 0,12 selanjutnya melakukan perhitungan QER, sebagai berikut :

$$score\ kunjung = \frac{|\ pkunjung + qkunjung\ |}{|\ pkunjung - qkunjung\ |}$$

$$score\ kunjung = \frac{|\ 0{,}208 + 0{,}12\ |}{|\ 0{,}208 - 0{,}12\ |} = 3{,}73$$

3.6 Klasifikasi : Support Vector Machine

Support Vector Machine (SVM) merupakan sistem untuk melakukan klasifikasi menggunakan ruang hipotesis berupa fungsi-fungsi linear dalam sebuah ruang fitur (feature space) berdimensi tinggi [20].Konsep SVM yaitu berusaha menemukan fungsi pemisah (hyperplane) terbaik diantara fungsi yang tidak terbatas jumlahnya. Hyperplane merupakan pemisah terbaik antara kedua kelas dapat ditemukan dengan mengukur margin hyperplane tersebut dan mencari titik maksimalnya. Dividing lines merupakan kombinasi dari vektor yang memberikan keputusan fungsi (kelas atau bukan kelas) untuk classifier SVM.

Penilaian kemudian dibuat dengan menilai *score* apakah positif atau negatif yang merepresentasikan di sisi mana dari garis pemisah dokumen berada. Sejauh ini fungsi kernel dapat diasumsikan sebagai dot product antar dua vector.

Prinsip kerja SVM yaitu mengklasifikasi secara linear (linear classifier), kemudian SVM dikembangkan dan dapat bekerja pada klasifikasi non linear, sehingga SVM memiliki dua formulasi optimasi yaitu klasifikasi linear dan klasifikasi non-linear. Dalam melakukan klasifikasi non-linear SVM diharuskan memodifikasi formula untuk menyelesaikan permasalahan tersebut karena tidak memiliki solusi lain yaitu dengan cara memasukkan fungsi Kernel.

Terdapat beberapa jenis kernel yang akan dirangkum pada tabel dibawah ini:

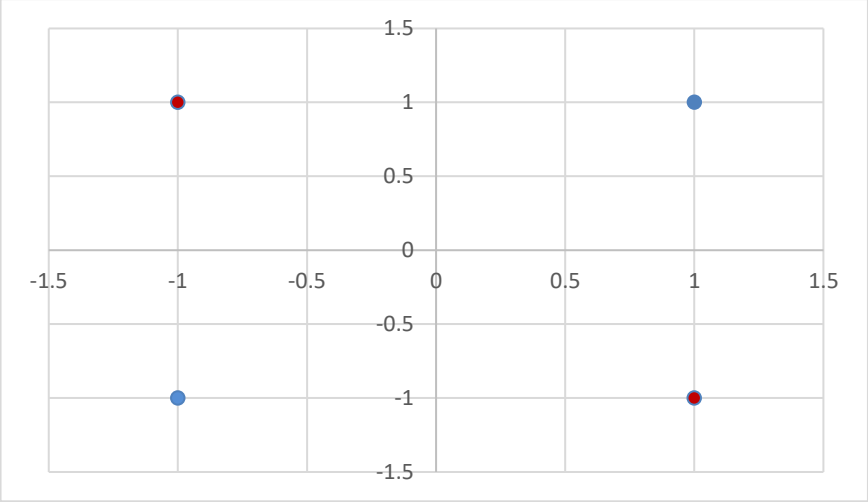
Table 11 Jenis Kernel

Jenis Kernel	Definisi
Polynomial	$K(x_i,x_j) = (x_i,x_j + 1)^p$
Gaussian RBF	$K(x_i, x_j) = \exp(-\frac{\|\|x_i - x_j\|\|^2}{2\sigma^2})$
Signoid	$K(x_i,x_j) = \tanh(\alpha x_i,x_j + \beta)$
Linear	$K(x_i,x_j) = x_i^t\, x_j$

Berdasarkan penjelasan diatas, penulis memberikan skema perhitungan dengan persamaan diatas. Seperti yang dilihat pada tebel dibawah ini menjelaskan ilustrasi dataset yang akan digunakan sebagai contoh skema perhitungan Support Vector Machine. Dataset ini memiliki nilai x sebagai inisialisasi dokumen dan y sebagai lebel kelas pada dokumen.

x1	x2	Kelas (Y)
1	1	1
1	-1	-1
-1	1	-1
-1	-1	1

Setelah mengetahui dataset yang akan digunakan maka langkah selanjutnya adalah melakukan visualisasi data untuk mengetahui bentuk kurva dengan jelas. Pada Gambar dibawah ini merupakan hasil dari visualisasi dataset dan memiliki 2 macam titik yaitu titik biru sebagai kelas positif dan titik merah sebagai kelas negative.



Pada gambar diatas terlihat bahwa kelompok data tidak linear, maka perlu menggunakan fungsi kernel. Formulasi yang digunakan untuk menyelesaikan permasalahan tersebut menggunakan fungsi kernel Polinomial ordo 2, dengan perhitungan sebagai berikut :

$$K(x,y) = (x.y + c)^d$$ dengan ketentuan c = 1 dan d = 2 .

- Fungsi kernel dituliskan kembali menjadi :

$$K(x,x_i) = (x^t.x_i + 1)^2 \quad \text{dengan } w = \sum_{i=1}^{n} \alpha_i y_i \Phi(x_i)$$

- Menghitung matriks kernel K

$$K(x_i, x_j) = \Phi(x_i).\Phi(x_j)$$

- Dimisalkan mengihitung K(M,N) dengan M = (1,1) dan N = (1,-1)

Maka : K(M=(1,1) , N=(1,-1))

$= \left(\left((M_1.N_1) + (M_2.N_2)\right) + 1\right)^2$

$= ((M_1.N_1) + (M_2.N_2))^2 + 2((M_1.N_1) + (M_2.N_2)) \cdot 1 + 1^2$

$= (M_1.N_1)^2 + 2(M_1.N_1)(M_2.N_2) + (M_2.N_2)^2 + 2(M_1.N_1) + 2(M_2.N_2) + 1$

$$= \begin{bmatrix} (M_1)^2 \\ \sqrt{2M_1M_2} \\ (M_2)^2 \\ \sqrt{2M_1} \\ \sqrt{2}M_2 \\ 1 \end{bmatrix} \begin{bmatrix} (N_1)^2 \\ \sqrt{2N_1N_2} \\ (N_2)^2 \\ \sqrt{2N_1} \\ \sqrt{2}N_2 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1^2 \\ \sqrt{2.1.1} \\ 1^2 \\ \sqrt{2.1} \\ \sqrt{2}.1 \\ 1 \end{bmatrix} \begin{bmatrix} 1^2 \\ \sqrt{2.1.-1} \\ -1^2 \\ \sqrt{2.1} \\ \sqrt{2}.-1 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 \\ \sqrt{2} \\ 1 \\ \sqrt{2} \\ \sqrt{2} \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ -\sqrt{2} \\ 1 \\ \sqrt{2} \\ -\sqrt{2} \\ 1 \end{bmatrix} = 1 - 2 + 1 + 2 - 2 + 1 = 1$$

Maka hasil semua matrik nya adalah :

x_1	x_1	$K(1,1) = (x_1 . x_1 + 1)^2 = (1 . 1 + 1 . 1 + 1)^2 = 9$
	x_2	$K(1,2) = (x_1 . x_2 + 1)^2 = (1 . 1 + 1 . (-1) + 1)^2 = 1$
	x_3	$K(1,3) = (x_1 . x_3 + 1)^2 = (1 . (-1) + 1 . 1 + 1)^2 = 1$
	x_4	$K(1,4) = (x_1 . x_4 + 1)^2 = (1 . (-1) + 1 . (-1) + 1)^2 = 1$
x_2	x_1	$K(2,1) = (x_2 . x_1 + 1)^2 = (1 . 1 + (-1) . 1 + 1)^2 = 1$
	x_2	$K(2,2) = (x_2 . x_2 + 1)^2 = (1 . 1 + (-1) . (-1) + 1)^2 = 9$
	x_3	$K(2,3) = (x_2 . x_3 + 1)^2 = (1 . (-1) + (-1) . 1 + 1)^2 = 1$
	x_4	$K(2,4) = (x_2 . x_4 + 1)^2 = (1 . (-1) + (-1) . (-1) + 1)^2 = 1$
x_3	x_1	$K(3,1) = (x_3 . x_1 + 1)^2 = ((-1) . 1 + 1 . 1 + 1)^2 = 1$
	x_2	$K(3,2) = (x_3 . x_2 + 1)^2 = ((-1) . 1 + 1 . (-1) + 1)^2 = 1$
	x_3	$K(3,3) = (x_3 . x_3 + 1)^2 = ((-1) . (-1) + 1 . 1 + 1)^2 = 9$
	x_4	$K(3,4) = (x_3 . x_4 + 1)^2 = ((-1) . 1 + 1 . (-1) + 1)^2 = 1$
x_4	x_1	$K(4,1) = (x_4 . x_1 + 1)^2 = ((-1) . 1 + (-1) . 1 + 1)^2 = 1$
	x_2	$K(4,2) = (x_4 . x_2 + 1)^2 = ((-1) . 1 + (-1) . (-1) + 1)^2 = 1$
	x_3	$K(4,3) = (x_4 . x_3 + 1)^2 = ((-1) . (-1) + (-1) . 1 + 1)^2 = 1$
	x_4	$K(4,4) = (x_4 . x_4 + 1)^2 = ((-1) . (-1) + (-1) . (-1) + 1)^2 = 9$

Didapatkan matrix kernel K dengan ukuran N x N :

$$K(x,x_i) = \begin{bmatrix} 9 & 1 & 1 & 1 \\ 1 & 9 & 1 & 1 \\ 1 & 1 & 9 & 1 \\ 1 & 1 & 1 & 9 \end{bmatrix}$$

- Setiap elemen matrix kernel $K(x,x_i)$ digunakan untuk menggantikan dot-product x_i,x_j dalam persamaan dualitas Lagrange Multiplier.

Maksimalkan Ld : $\sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j (x_i . x_j)$

Dimisalkan didapatkan nilai Max Ld dengan $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.125$ sehingga Ld = 0.25

- Hitung nilai w dan b

$$w = \sum_{i=1}^{N} \alpha_i y_i \emptyset(X_i)$$

$$w = \sum_{i=1}^{4} \alpha_i y_i \emptyset(X_i) = \alpha_1 y_1 \emptyset(X_1) + \alpha_2 y_2 \emptyset(X_2) + \alpha_3 y_3 \emptyset(X_3) + \alpha_4 y_4 \emptyset(X_4)$$

$$w = (-1)0.125\begin{bmatrix} x_1^{1^2} = 1^2 = 1 \\ \sqrt{2x_1^1 x_2^1} = \sqrt{2(1)(1)} = \sqrt{2} \\ x_1^{1^2} = 1^2 = 1 \\ \sqrt{2x_1^1} = \sqrt{2(1)} = \sqrt{2} \\ \sqrt{2x_2^1} = \sqrt{2(1)} = \sqrt{2} \\ 1 \end{bmatrix} + 0.125\begin{bmatrix} 1 \\ -\sqrt{2} \\ 1 \\ \sqrt{2} \\ -\sqrt{2} \\ 1 \end{bmatrix} + 0.125\begin{bmatrix} 1 \\ -\sqrt{2} \\ 1 \\ -\sqrt{2} \\ \sqrt{2} \\ 1 \end{bmatrix}$$

$$- 0.125\begin{bmatrix} 1 \\ \sqrt{2} \\ 1 \\ -\sqrt{2} \\ -\sqrt{2} \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ -0{,}71 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Pilih salah satu Support Vector dari kelas "+1" dan "-1" untuk menghitung nilai b

$$b = -\frac{1}{2}\,(w\,.\,x^{+} + w.w^{-})$$

$$= -\frac{1}{2}\left[\begin{bmatrix} 0 \\ -0{,}71 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}\begin{bmatrix} 1 \\ -\sqrt{2} \\ 1 \\ \sqrt{2} \\ -\sqrt{2} \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ -0{,}71 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}\begin{bmatrix} 1 \\ \sqrt{2} \\ 1 \\ \sqrt{2} \\ \sqrt{2} \\ 1 \end{bmatrix}\right]$$

$$= -\frac{1}{2}\left((-0{,}71)(-\sqrt{2}) + (0{,}71)(\sqrt{2})\right) = 0$$

- Maka model SVM siap digunakan untuk proses klasifikasi.

$$f(\emptyset(x)) = sign\,(w\,.\,\emptyset(x) + b) = sign\left(\sum_{i=1}^{N} \alpha_i\, y_i \emptyset(X_i.\emptyset(x) + b\right)$$

Misalkan data uji/data tes x_t = (1,5) maka $K(x_i, x_t) = \Phi(x_i).\Phi(x_t)$

x_t = (1,5)	x_1	K(1,1) = $(x_1\,.\,x_t + 1)^2$ = (1 . 1 + 5 . 1 + 1)^2 = 49
	x_2	K(1,2) = $(x_1\,.\,x_t + 1)^2$ = (1 . 1 + 5 . (-1) + 1)^2 = 9
	x_3	K(1,3) = $(x_1\,.\,x_t + 1)^2$ = (1 . (-1) + 5 . 1 + 1)^2 = 25
	x_4	K(1,4) = $(x_1\,.\,x_t + 1)^2$ = (1 . (-1) + 5 . (-1) + 1)^2 = 25

$$f(\emptyset(x_t)) = sign\,(w\,.\,\emptyset((1{,}5)) + b) = sign\,(-6.125 + 1.125 + 3.125 - 3.125 + 0)$$
$$= sign(-5) = -1$$

Jadi, data x_t = (1,5) tersebut masuk pada kelas negative.

3.7 Confusion Matrix

Untuk menghitung nilai akurasi, precision, recall, dan FI *Score* maka pada penelitian ini menggunakan confusion matriks sebagai evaluasi performansi. Akurasi merupakan perbandingan antara data yang terklasifikasi benar dengan keseluruhan data yang ada, precision merupakan jumlah data kategori positif yang diklasifikasikan secara benar dibagi total data yang diklasifikasikan positif. Recall merupakan berapa persentasi data kategori positif yang terklasifikasi dengan benar oleh sistem [21]. F1-Score digunakan untuk membantu mengukur nilai recall dan presisi secara bersamaan sehingga bisa dilihat pada implementasinya secara keseluruhan.

Pada table 12 dibawah ini akan dijelaskan mengenai confusion matrix:

Table 12 Confusion Matrix

		Predicted Class	
		Positive	***Negative***
Actual Class	***Positive***	*True Positive (TP)*	*False Negative (FN)*
	Negative	*False Positive (FP)*	*True Negative (TN)*

$$Accuracy = \frac{TP+TN}{(TP+TN+FP+FN)} \times 100\% \quad (10)$$

$$\text{Precision} = \frac{TP}{(TP+FP)} \times 100\% \quad (11)$$

$$Recall\ \frac{TP}{(TP+FN)} \times 100\% \quad (12)$$

$$F1\ Score = 2 \times \frac{(Precision \times Recall)}{Precision+Recall} \quad (13)$$

Tabel 12 menunjukkan kombinasi nilai kelas prediksi dan kelas actual yang akan digunakan untuk menghitung nilai akurasi, precision, recall, dan F1 Score. True Positive (TP) merupakan sebuah kondisi ketika kelas prediksi bermakna positif dan kelas aktualnya benar. True Negative (TN) merupakan sebuah kondisi ketika kelas prediksi bermakna negative dan kelas aktualnya benar. False Positive (FP) merupakan sebuah kondisi ketika kelas prediksi bermakna positif dan kelas aktualnya salah. False Negative (FN) merupakan sebuah kondisi ketika kelas prediksi bermakna negative dan kelas aktualnya salah.

4. Evaluasi

Pada proses pengujian terdiri dari 2 skenario pengujian. Scenario yang pertama adalah pengujian perbandingan kernel pada SVM. Scenario kedua adalah pengujian perbandingan variasi rasio pada seleksi fitur QER. Scenario yang dilakukan tersebut menggunakan keadaan pembagian data latih dan data uji sebesar 80:20

4.1 Pengujian Perbandingan Kernel pada Support Vector Machine

Pengujian ini dilakukan pada dataset yang sudah melalui proses preprocessing, pembobotan TF-IDF dan ekstraksi fitur Bigram. Klasifikasi ini menggunakan seleksi fitur QER dan algoritma SVM untuk mengetahui penggunaan kernel yang terbaik. Pada pengujian ini kernel-kernel yang akan digunakan yaitu linear, polynomial dan RBF. Ketiga kernel tersebut memiliki peranan penting untuk mengklasifikasi dataset. Pada setiap dataset yang akan diuji penggunaan parameter juga dibutuhkan untuk mengetahui hasil akurasi terbaik.

4.1.1 Kernel Linear

Kernel Linear merupakan fungsi kernel yang digunakan ketika data sudah terpisah secara linear. Dalam melakukan pengujian menggunakan fungsi kernel Linear perlu dilakukan optimasi parameter C (cost), pengoptimalan parameter C dapat dilakukan dengan cara trial and error [22]. Tabel dibawah ini merupakan hasil penentuan parameter C terbaik untuk penggunaan kernel Linear.

Table 13 Nilai akurasi parameter terbaik kernel Linear

Parameter	Akurasi Kernel Linear
C = 0.001	49%
C = 0.01	49%
C = 0.1	94%
C = 1	98%
C = 10	98%
C = 100	98%

Dari tabel 13 hasil yang diperoleh menunjukkan parameter C terbaik adalah C = 1 dengan hasil akurasi yang diperoleh adalah 98%. Perolehan nilai Accuracy yang didapatkan pada pengujian menggunakan Cross Validation dengan nilai CV = 5. Selanjutnya dari hasil pemilihan parameter terbaik diatas dapat dibuat confusion matrix sebagai metode evaluasi untuk mengetahui hasil klasifikasi.

Table 14 Confusion Matrix Kernel Linear

Accuracy	98%	
F1-Score	98%	
	Positive	Negative
Pred. Positive	4	87
Pred. Negative	86	0

Pada tabel 14 menjelaskan hasil Confusion Matrix menggunakan kernel Linear mendapatkan nilai prediksi dengan masing-masing kelas. Nilai prediksi tersebut merupakan hasil record data test yang berukuran 20% dari dataset. Hasil yang diperoleh adalah F1-Score memiliki nilai sebesar 98% dan akurasi sebesar 98%.

4.1.2 Kernel Polynomial

Kernel Polynomial merupakan fungsi kernel Non-Linear yang cocok digunakan dalam kondisi semua training dataset nya telah di normalisasi. Parameter yang dibutuhkan untuk menggunakan kernel ini adalah C(cost) dan d (degree). Pengujian ini juga menggunakan cara trial and error seperti sebelumnya untuk pengoptimalan parameter C dan d terbaik sehinggal menghasilkan seperti tabel dibawah ini.

Table 15 Nilai akurasi parameter terbaik kernel Polinomial

Parameter C = 1	Akurasi Kernel Polinomial
d = 1	98%
d = 2	100%
d = 3	100%

Dari keseluruhan trial and error yang dilakukan untuk memperoleh hasil akurasi terbaik dalam pengujian parameter d, hasil yang didapatkan saat menggunakan parameter d = 1 adalah 98% , d =2 adalah 100% dan

d = 3 adalah 100% . Sehingga dapat dilihat bahwa penggunaan parameter d = 2 adalah parameter terbaik dalam pengujian ini. Perolehan nilai Accuracy yang didapatkan pada pengujian menggunakan Cross Validation dengan nilai CV = 5. Selanjutnya dari hasil pemilihan parameter terbaik diatas dapat dibuat confusion matrix sebagai metode evaluasi untuk mengetahui hasil klasifikasi.

Table 16 Confusion Matrix Kernel Polinomial

Accuracy	100%	
F1-Score	100%	
	Positive	Negative
Pred. Positive	0	91
Pred. Negative	86	0

Pada tabel 16 menjelaskan hasil Confusion Matrix menggunakan kernel Polinomial mendapatkan nilai prediksi dengan masing-masing kelas. Nilai prediksi tersebut merupakan hasil record data test yang berukuran 20% dari dataset. Hasil yang diperoleh adalah F1-Score memiliki nilai sebesar 100% dan akurasi sebesar 100%.

4.1.3 Kernel RBF

Kernel RBF (Radial Basis Function) merupakan fungsi kernel yang digunakan dalam analisis ketika data tidak terpisah secara Linear. Parameter yang dibutuhkan untuk menggunakan kernel ini adalah C (cost) dan γ (gamma). Seperti pada pengujian fungsi kernel sebelumnya, pengujian ini menggunakan cara trial and error untuk pengomptimalan parameter C dan γ sehingga menghasilkan seperti tabel dibawah ini.

Table 17 Nilai akurasi parameter terbaik kernel RBF

Parameter C = 1	Akurasi Kernel RBF
γ = 1	100%
γ = 2	100%
γ = 3	100%
γ = 4	100%

Dari keseluruhan trial and error yang dilakukan untuk memperoleh hasil akurasi terbaik dalam pengujian parameter γ, hasil yang didapatkan saat menggunakan parameter γ = 1 hingga γ = 4 memiliki hasil akurasi sebesar 100%. Sehingga dapat dilihat bahwa penggunaan parameter γ = 1 sudah dapat digunakan dalam pengujian kernel RBF. Perolehan nilai Accuracy yang didapatkan pada pengujian menggunakan Cross Validation dengan nilai CV = 5. Selanjutnya dari hasil pemilihan parameter terbaik diatas dapat dibuat confusion matrix sebagai metode evaluasi untuk mengetahui hasil klasifikasi.

Table 18 Confusion Matrix Kernel RBF

Accuracy	100%	
F1-Score	100%	
	Positive	Negative
Pred. Positive	0	91
Pred. Negative	86	0

Pada tabel 18 menjelaskan hasil Confusion Matrix menggunakan kernel RBF mendapatkan nilai prediksi dengan masing-masing kelas. Nilai prediksi tersebut merupakan hasil record data test yang berukuran 20% dari dataset. Hasil yang diperoleh adalah F1-Score memiliki nilai sebesar 100% dan akurasi sebesar 100%.

4.1.4 Perbandingan Pengujian Fungsi Kernel

Setelah melakukan pengujian 3 fungsi kernel menggunakan kernel Linear, Polinomial, dan RBF dari data ulasan pariwisata Pulau Madura, kemudian akan ditentukan manakah fungsi kernel yang paling sesuai dalam klasifikasi, yang akan dirangkum pada tabel rangkuman dibawah ini dengan masing-masing nilai akurasi yang diperoleh dari ketiga fungsi kernel, menggunakan Cross Validation dengan CV = 10

Table 19 Pengujian Setiap Kernel pada SVM

Nilai Akurasi		
Kernel Linear	**Kernel Polinomial**	**Kernel RBF**
0,972	1,00	1,00
0,956	0,972	0,972
0,944	1,00	1,00
0,972	1,00	1,00
0,944	0,972	0,972
0,986	1,00	0,986

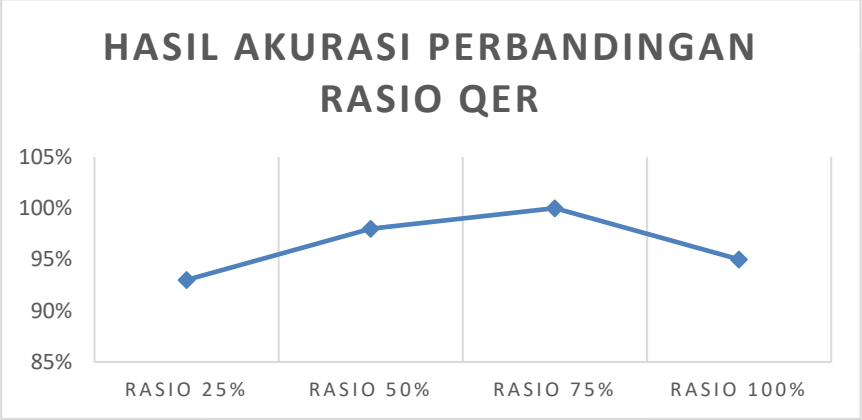
0,971	1,00	1,00
0,957	1,00	1,00
0,986	1,00	1,00
0,944	1,00	1,00
Mean = 0,9636 Best = 0,986	Mean = 0,9944 Best = 1,00	Mean = 0,993 Best = 1,00

Setelah melakukan 10 kali pengujian dengan membandingkan setiap kernel pada Support Vector Machine, dapat dilihat pada tabel 19 untuk penggunaan kernel Polinomial dan kernel RBF memiliki nilai akurasi yang tinggi yaitu 100% dengan rata-rata 99% sedangkan penggunaan kernel Linear memiliki nilai akurasi 98% dengan rata-rata 96%.

Berdasarkan tabel 19 menunjukkan bahwa penggunaan fungsi kernel yang tepat untuk pengujian ini adalah fungsi kernel Polinomial dan kernel RBF dengan hasil akurasi sebesar 100%. Sehingga terlihat bahwa dataset yang digunakan pada penelitian ini merupakan data Non-Linear sehingga penggunaan kernel Polinomial dan kernel RBF sangat cocok untuk digunakan dalam penelitian ini. Selanjutnya penguji hanya mengambil 1 kernel yang akan digunakan untuk pengujian berikutnya, yaitu pengujian perbandingan Rasio pada Seleksi Fitur Query Expansion Ranking, kernel yang digunakan pada pengujian selanjutnya adalah kenel Polinomial.

4.2 Pengujian Perbandingan penghapusan fitur pada Query Expansion Ranking

Pengujian ini dilakukan pada dataset yang sudah melalui proses preprocessing, pembobotan TF-IDF, ekstraksi fitur Bigram dan menggunakan kernel Polinomial . Klasifikasi ini menggunakan seleksi fitur QER dan algoritma SVM untuk mengetahui penghapusan fitur berapa yang memiliki nilai akurasi terbaik. Pengujian penghapusan fitur ini menggunakan kondisi fitur yang dihapus dari 25%, 50%, 75% dan 100% dari keseluruhan fitur.



Gambar 2 Pengujian Rasio Feature Selection

Gambar 2 menunjukkan grafik hasil akurasi dari rasio QER untuk pengklasifikasian menggunakan SVM. Hasil akurasi saat menggunakan rasio 100% atau menggunakan seluruh fitur untuk klasifikasi memiliki hasil akurasi sebesar 95%, kemudian rasio diturunkan menjadi 75% dan memiliki peningkatan pada hasil akurasinya sebesar 100%. Rasio diturunkan lagi menjadi 50% hasil akurasi nya mengalami penurunan yaitu sebesar 98%, selanjutnya rasio diturunkan kembali menjadi 25% dan hasil akurasi kembali mengalami penurunan sebesar 93%.

Penggunaan rasio 100% memiliki nilai akurasi rendah dikarenakan semua fitur digunakan dan tidak semua memiliki relevansi pada dokumen yang diujikan. Penggunaan rasio 75% dan 50% dapat meningkatkan hasil akurasi dikarenakan fitur yang digunakan memiliki relevansi pada dokumen yang diujikan dan bukan merupakan fitur yang noise. Penggunaan rasio 25% mengalami penurunan dikarenakan pengurangan fitur yang terlalu banyak sehingga fitur-fitur yang seharusnya memiliki relevansi dengan dokumen yang diuji juga akan ikut terbuang. Berdasarkan hasil yang didapatkan pada pengujian ini, pengujian dengan penggunaan rasio 75% memiliki hasil akurasi yang tinggi yaitu sebesar 100%.

5. Kesimpulan dan Saran

Berdasarkan hasil pengujian dan analisis dari klasifikasi ulasan pariwisata di Pulau Madura menggunakan metode *Support Vector Machine* dan Query Expansion Ranking dapat disimpulkan sebagai berikut:

Penelitian ini memiliki 5 proses utama yaitu pengumpulan data, preprocessing, ekstraksi fitur dengan Bigram, pembobotan fitur dengan TF-IDF, kemudian melakukan seleksi fitur dengan Query Expansion Ranking dan melakukan klasifikasi dengan Support Vector Machine.

Pengujian pertama yaitu perbandingan kernel pada Support Vector Machine dengan parameter yang sudah di tentukan sebelumnya. Hasil akurasi untuk kernel Linear dengan menggunakan parameter C = 1 memiliki hasil 98%, pada kernel Polinomial dengan menggunakan parameter C = 1 dan degree = 2 memiliki hasil 100%, dan pada kernel RBF dengan parameter C = 1 dan Gamma = 1 memiliki hasil 100%. Sehingga pada pengujian pertama menghasilkan kernel terbaik untuk pengklasifikasian pada penelitian ini yaitu kernel Polinomial dan kernel RBF dengan hasil akurasi 100%.

Pengujian kedua yaitu membandingkan rasio pada seleksi fitur Query Expansion Ranking dengan menggunakan perbanding rasio 25%, 50%, 75%, dan 100%. Hasil akurasi untuk perbandingan rasio QER pada rasio 25% memiliki hasil 93%, rasio 50% memiliki hasil 98%, untuk rasio 75% memiliki hasil 100% dan pada rasio 100% memiliki hasil 95%. Sehingga pada pengujian kedua ini menghasilkan rasio terbaik untuk penelitian ini yaitu pada penggunaan rasio 75% dengan hasil 100%.

Ada beberapa saran agar penelitian selanjutnya dapat lebih baik lagi yaitu :

1. Menambahkan data agar lebih banyak dan lebih bervariatif
2. Melakukan perbandingan pada seleksi fitur untuk mengetahui seleksi fitur mana yang lebih baik

Referensi

[1] D. Indonesia, Digital 2020: Indonesia — DataReportal – Global Digital Insights, Datareportal. (2020) 1–92. https://datareportal.com/reports/digital-2020-indonesia.

[2] F. FANANI, I. Alfi Bustoni, Klasifikasi Review Software Pada Google Play Menggunakan Pendekatan Analisis Sentimen, Ilmu Komun. (2017).

[3] W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Eng. J. 5 (2014) 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011.

[4] M.J. Ubaidillah, I. Munadhif, N. Rinanto, Klasifikasi Gelombang Otot Lengan Pada Robot Manipulator Menggunakan Support Vector Machine, Rekayasa. 12 (2019) 91–97. https://doi.org/10.21107/rekayasa.v12i2.5406.

[5] B.P. Putra, B. Irawan, C. Setianingsih, F.T. Elektro, U. Telkom, D. Learning, Deteksi Ujaran Kebencian Dengan Menggunakan Algoritma Convolutional Neural Network Pada Gambar Hatespeech Detection Using Convolutional Neural Network Algorithm Based on Image, 5 (2018) 2395–2402.

[6] A. Deny Nusyirwan, Jurnal Ilmiah Pendidikan Teknik Kejuruan (JIPTEK), J. Ilm. Pendidik. Tek. Kejuru. 101 (2019) https://jurnal.uns.ac.id/jptk.

[7] T. Parlar, S.A. Ozel, A new feature selection method for sentiment analysis of Turkish reviews, Proc. 2016 Int. Symp. Innov. Intell. Syst. Appl. INISTA 2016. (2016). https://doi.org/10.1109/INISTA.2016.7571833.

[8] S. Fanissa, M.A. Fauzi, S. Adinugroho, Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking, (2018).

[9] I.T.S.A. Pamungkas, Analisis Sentimen Terhadap Tokoh Publik Menggunakan Algoritma Support Vector Machine (Svm), Log!K@. 8 (2018) 69–79.

[10] E. Indrayuni, Analisa Sentimen Review Hotel Menggunakan Algoritma Support Vector Machine Berbasis Particle Swarm Optimization, J. Evolusi Vol. 4 Nomor 2 - 2016. 4 (2016) 20–27.

[11] andi nurul Hidayat, Analisis Sentimen Terhadap Wacana Politik Pada Media Masa Online Menggunakan Algoritma Support Vector Machine Dan Naive Bayes, J. Elektron. Sistim Inf. Dan Komput. 1 (2015) 1–7.

[12] S.K. Lidya, O.S. Sitompul, S. Efendi, Sentiment Analysis Pada Teks Bahasa Indonesia Menggunakan Support Vector Machine (Svm), Semin. Nas. Teknol. Dan Komun. 2015. 2015 (2015) 1–8. https://doi.org/10.1016/j.eswa.2013.08.047.

[13] V. Narayanan, I. Arora, A. Bhatia, Fast and accurate sentiment classification using an enhanced Naive Bayes model, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 8206 LNCS (2013) 194–201. https://doi.org/10.1007/978-3-642-41278-3_24.

[14] N.D. Mentari, M.A. Fauzi, L. Muflikhah, Analisis Sentimen Kurikulum 2013 Pada Sosial Media Twitter Menggunakan Metode K-Nearest Neighbor dan Feature Selection Query Expansion Ranking, J. Pengemb. Teknol. Inf. Dan Ilmu Komput. Univ. Brawijaya. 2 (2018) 2739–2743.

[15] M.C. Untoro, J.L. Buliali, Penanganan imbalance class data laboratorium kesehatan dengan majority weighted minority oversampling technique, Regist. J. Ilm. Teknol. Sist. Inf. 4 (2018) 23–29. https://doi.org/10.26594/register.v4i1.1184.

[16] E. Nour, IMPLEMENTASI METODE CONVOLUTIONAL NEURAL NETWORK UNTUK KLASIFIKASI TANAMAN PADA CITRA RESOLUSI TINGGI (The Implementation of Convolutional Neural Network Method for Agricultural Plant Classification in High Resolution Imagery), (2018) 61–68.

[17] H. Widayanto, A.F.H. Huda, Comparison Nazief Adriani And CS Stemmer Algorithm For Stemm Real Data, E-Proceeding Eng. 4 (2017) 5215–5222.

[18] H. MacLaughlin, S. Greenwood, Weight management of obese patients on the renal ward, J. Ren. Nurs. 2 (2010) 116–121. https://doi.org/10.12968/jorn.2010.2.3.48079.

[19] B.A.B. Ii, L. Teori, D.A.N. Studi, BAB II LANDASAN TEORI DAN STUDI LITERATUR 2.1. Definisi Manajemen Proyek Menurut, (2004) 22–59.

[20] B. Santosa, Data mining:Teknik Pemanfaatan Data untuk Keperluan Bisnis. Yogyakarta: Graha Ilmu-Bisnis.Edisi Pertama., Data MiningTeknik Pemanfaat. Data Untuk Keperluan Bisnis. Yogyakarta Graha Ilmu- Bisnis.Edisi Pertama. 33 (2007) 365–373.

[21] F.M. Alfath, I. Asror, Y.R. Murti, Klasifikasi Emosi pada Tweet di Twitter Menggunakan Metode K-

Nearest Neighbor, (n.d.).

[22] S.N.D. Pratiwi, B.S.S. Ulama, Klasifikasi Email Spam dengan Menggunakan Metode Support Vector Machine dan k-Nearest Neighbor, J. Sains Dan Seni ITS. 5 (2016) 344–349.