SENTIMENT ANALYSIS KOMENTAR YOUTUBE SAMSUNG S20 MENGGUNAKAN METODE MAJORITY VOTE

Risang Putra Pradana[1], Deni Arifianto[2], Habibatul Azizah Al Faruq[3]

Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Jember

Jln. Karimata No. 49 Jember Kode Pos 68121

email: risangard@gmail.com

ABSTRAK

Samsung S20 merupakan flagship keluar Samsung, meskipun Samsung Galaxy S20 hadir dengan banyak fitur, namun informasi tentang tingkat penerimaan konsumen terhadap produk ini tetap dibutuhkan. Dalam penelitian ini, penulis akan melakukan analisis sentimen berdasarkan komentar yang terdapat dalam YouTube review Samsung S20, dengan menerapkan proses text mining Majority vote yang terdiri dari perbandingan metode Support Vector Machine, K-Nearest Neighbors, dan Naïve Bayes, untuk mengklasifikasikan apakah teks termasuk dalam sentimen positif, negatif dan netral. Data diklasifikasikan secara manual dengan mengelompokkan menjadi kelas sentimen positif, negatif dan netral kemudian secara otomatis data latih akan mengambil beberapa data untuk data uji dan menguji kemampuan sistem. Hasil yang didapatkan dari penelitian ini adalah dengan menggunakan Majority vote mendapatkan nilai akurasi 0.8 dengan menggunakan 100 data dan mendapatkan nilai akurasi 0.48 dengan menggunakan 50 data uji.

Kata Kunci : Samsung S20, YouTube, text mining, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes, Majority Vote

ABSTRACT

Samsung S20 is the latest Samsung's flagship, even though Samsung Galaxy S20 comes with many features, information about the level of consumer acceptance of this product is necessary. In this study, the author will conduct a sentiment analysis based on the customer comments in the YouTube review of the Samsung S20, by applying the Majority vote text mining process which consists of a comparison of the Support Vector Machine, K-Nearest Neighbors, and Naïve Bayes methods, to classify whether the text is included in the positive, negative and neutral sentiment. Data is classified manually by grouping into positive, negative and neutral sentiment classes then automatically the accustomed data will take some data for test data and test the system's ability. The results obtained from this study are the data accuracy from 100 data is 0.8 and 0.48 from 50 data tests by using Majority Vote.

Keywords: Samsung S20, YouTube, text mining, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes, Majority Vote

1. PENDAHULUAN

Samsung Galaxy S20 adalah smartphone flagship keluaran terbaru Samsung pada tahun 2020. Memiliki ukuran yang hanya 6.2-inch membuatnya mudah digunakan dengan satu tangan, ringkas dan mudah dibawa. Tampilan dari Samsung Galaxy S20 terkesan simple dan elegan karena bagian layar dan belakang pada Samsung Galaxy S20 terbuat dari kaca. Menggunakan layar Dynamic Amoled dengan fitur 120hz pastinya membuat mata nyaman saat memandang layar Samsung Galaxy S20. Samsung Galaxy S20 memiliki speaker stereo yang lantang dan jelas, membuat penggunanya akan nyaman ketika mendengarkan music. Dalam penggunaan sehari-hari dengan mengandalkan prosesor Exynos 990, RAM 8GB dan media penyimpanan internal sebesar 128 GB membuat Samsung Galaxy S20 dapat digunakan dalam berbagai hal, seperti memainkan game, menjalankan berbagai jenis multimedia, browsing, atau membuka Instagram. Berkat baterai sebesar 4000 mAh, Samsung Galaxy S20 dapat digunakan seharian penuh. Dan fitur andalan Samsung Galaxy S20 adalah kamera. Samsung Galaxy S20 memiliki tiga buah kamera yaitu kamera Ultra-wide 12 MP, kamera Wide 12 MP dan kamera Telephoto 64 MP, dengan tiga buah kamera yang ada, dipastikan kamera Samsung Galaxy S20 dapat diandalkan dalam berbagai aplikasi photography, seperti untuk mengambil foto pemandangan, mengambil foto pada malam hari atau sekedar mengambil foto untuk di upload di Instagram.

Meskipun Samsung Galaxy S20 hadir dengan banyak fitur, namun informasi tentang tingkat penerimaan konsumen terhadap produk ini tetap dibutuhkan. Informasi itu disebut Consumer Preference. Consumer Preference adalah penilaian subjektif konsumen individu diukur dari kepuasan mereka dalam menggunakan produk yang telah mereka beli. Consumer Preference penting diketahui oleh produsen dan konsumen potensial. Konsumen potensial adalah konsumen yang berminat melakukan pembelian terhadap produk yang ditawarkan. Berdasarkan penelitian terdahulu (Næs et al., 2010; Rintyarna, 2016), Consumer preference umumnya di ekstrak dengan metode Conjoint Analysis. Conjoint Analysis adalah metode riset pemasaran yang paling banyak digunakan untuk menganalisis preferensi konsumen. Data yang digunakan pada teknik ini adalah data yang dikumpulkan dengan survei.

Teknik survei dianggap time-consuming dan costly (perlu waktu lama dan biaya yang besar). Penelitian Tugas Akhir ini mengusulkan sebuah pendekatan berbasis Teknik Sentiment analysis untuk mengetahui tingkat Consumer Preference tanpa melakukan survei. Analisis sentimen adalah bidang studi yang menganalisis pendapat seseorang, opini, evaluasi, penilaian, sikap dan emosi terhadap entitas seperti produk, layanan, organisasi, individu, masalah, peristiwa, topik dan atribut (Balya, 2019). Teknik Sentiment analysis akan diaplikasikan pada data komentar YouTube yang berkaitan dengan produk Samsung Galaxy S20.

Banyak macam metode untuk menghitung Sentiment analysis seperti SVM, KNN dan Naïve Bayes. Metode Majority vote yang digunakan oleh (Zamahsyari & Nurwidyantoro, 2017) dalam menghitung sentimen komentar dapat menghasilkan nilai presisi dan akurasi yang lebih baik yaitu sebesar 75% dan 72%, sedangkan metode KNN yang digunakan oleh (Rosdiansyah, 2014) menghasilkan nilai akurasi 70%, dan Naïve Bayes berdasarkan penelitian yang telah dilakukan oleh (Sipayung et al., 2016) menunjukkan Naïve Bayes dengan membagi perhitungan dengan dua bagian, mendapatkan hasil akurasi sistem memiliki nilai accuration kategori 77.14% dan untuk precision sentiment 99.12%, recall sentiment 72.9%, dan accuration sentiment 75.42%.

Untuk menyeimbangkan kelemahan dari masing – masing metode terdapat metode yang dapat membandingkan ketiga metode tersebut yaitu metode Sentiment analysis yang akan digunakan berbasis teknik Majority Vote. Majority vote adalah teknik membandingkan hasil klasifikasi dari sejumlah algoritma Machine Learning. Menurut Zamahsyari & Nurwidyantoro (2017) berpendapat teknik Majority vote mampu meningkatkan kinerja algoritma Machine Learning.

Penelitian dengan topik Sentiment analysis pada data komentar YouTube pernah dilakukan oleh (Balya, 2019) menggunakan Gaussian Naïve Bayes dan Multinominal Gaussian Naïve Bayes. Penelitian ini salah satunya dimaksudkan untuk mengetahui tingkat akurasi dari teknik Majority Vote. Algoritma yang akan digunakan dalam perbandingan dalam Majority vote adalah SVM, KNN dan Naïve Bayes.

Data yang akan digunakan dalam penelitian ini berasal dari channel YouTube "GSMArena Official". Channel tersebut dibuat pada tanggal 6 Februari 2007, di channel itu berisi tentang video hands-on, unboxing dan berbagai video yang terkait dengan dunia teknologi smartphone.

2. TINJAUAN PUSTAKA

2.1. Sentiment Analysis

Sentiment analysis adalah bidang studi yang menganalisis opini orang, opini, valuasi, penilaian, sikap dan emosi terhadap produk, layanan, organisasi, individu, masalah, peristiwa, topik, dan atribut (Balya, 2019). Dengan menggunakan *sentiment analysis*, dapat mengetahui tingkat *Consumer Preference* tanpa melakukan survei.

Sentimen menurut Kamus Besar Bahasa Indonesia (2016) adalah pendapat atau pandangan yang didasarkan pada perasaan yang berlebih – lebihan terhadap sesuatu (bertentangan dengan pertimbanagan pikiran). Opini menurut Kamus Besar Bahasa Indonesia (2016) adalah pendapat atau pikiran atau pendirian.

2.2. Penelitian Terkait

Penelitian (Sipayung et al., 2016) yang berjudul "Perancangan Sistem Analisis Sentimen Komentar Pelanggan Menggunakan Metode Naive Bayes Classifier" data diambil dari komentar pelanggan agoda.com. Tujuan dari penelitian ini untuk membantu pihak hotel dalam mengetahui dan mengelompokkan komentar pelanggan, berdasarkan kelompok kategori dan sentimen. Metode yang digunakan adalah metode *Naïve Bayes* Classification (NBC). Dalam penelitian ini dapat disimpulkan bahwa dengan menggunakan 175 data dengan menggunakan metode NBC mendapatkan nilai accuration kategori 77.14% dan untuk *precision sentiment* 99.12%, *recall sentiment* 72.9%, dan *accuration sentiment* 75.42%, sehingga sistem mampu mengembalikan dokumen dan kecocokan data yang tinggi.

Penelitian sejenis pernah diusulkan oleh Balya (2019) yang berjudul "Analisis Sentimen Pengguna YouTube di Indonesia pada Review Smartphone Menggunakan *Naïve Bayes*". Data berasal dari komentar YouTube yang berada di salah satu video *review channel* YouTube GadgetIn. Tujuan dari penelitian ini untuk mengetahui tingkat *sentiment positive*, *negative* dan netral terhadap komentar aplikasi YouTube yang terdapat pada video *review smartphone* secara otomatis, cepat dan tepat. Metode yang digunakan pada penelitian ini adalah algoritma *Naïve Bayes*. Dalam penelitian ini dapat disimpulkan bahwa menggunakan *Gaussian Naïve Bayes* menghasilkan akurasi sebesar 73% sementara dengan menggunakan *Multinomial Naïve Bayes* menghasilkan akurasi sebesar 81%, yang menyimpulkan bahwa *Multinomial Naïve Bayes* menghasilkan akurasi lebih baik dibandingkan dengan menggunakan Gaussian *Naïve Bayes*.

Ernawati & Wati (2018) pernah melakukan penelitian yang berjudul "Penerapan Algoritma *K-Nearest Neighbors*s Pada Analisis Sentimen Review Agen Travel". Data yang digunakan didapatkan melalui situs "Truspilot", data terdiri dari 100 *review* positif dan 100 *review* negatif. Tujuan dari penelitian ini adalah untuk mengklasifikasikan opini secara akurat menggunakan metode *K-Nearest Neighbors (KNN*). Dalam penelitian ini dapat disimpulkan bahwa dalam mengklasifikasi *K-Nearest Neighbors*, menggunakan 100 *review* positif dan 100 *review* negatif serta enam kata yang berhubungan dengan sentimen yaitu *Fast*, *Good*, *Great*, *Bad, Cancel* dan *Wait*. Nilai akurasi yang dihasilkan mencapai 87.00% dengan nilai AUC sebesar 0.916 masuk ke dalam kelompok *Excellent Classification*.

Sentiment analysis pada data artikel pernah dilakukan oleh Zamahsyari & Nurwidyantoro (2017). Data yang digunakan berasal dari 7500 artikel berita yang dibagi menjadi 15 kategori, setiap kategori terdiri dari 500 artikel. Tujuan penelitian ini adalah menganalisis berita dengan mudah untuk membantu pemerintah membuat kebijakan ekonomi. Dalam penelitian ini menyimpulkan bahwa membandingkan *decision tree*, *Random Forest* dan *Support Vector Machine* dengan menggunakan *Majority vote* performa yang dihasilkan dalam hal presisi dan akurasi mendapatkan hasil yang lebih baik.

Penelitian Tanesab dkk. (2017) yang berjudul "*Sentiment analysis Model Based On YouTube Comment Using Support Vector Machine*". Data yang digunakan berasal dari komentar di media sosial YouTube yang membahas tentang kinerja Ahok. Dalam penelitian ini dapat disimpulkan bahwa dengan menggunakan metode *Support Vector Machine(SVM)* dan menggunakan *Lexicon Based* dan *Confusion Matrix* untuk mengetahui hasil persentase bobot analisis terhadap *SVM*. Mendapatkan nilai *true positive* sebesar 91.1%.

Hustinawaty dkk. (2019) dalam menggunakan *KNN* untuk *sentiment analysis* pada data ulasan google tentang Pasar Lama Tangerang yang berjumlah 120 data. Tujuan dari penelitian ini adalah untuk memberikan wawasan kepada konsumen baru tentang Pasar Lama Tangerang berdasarkan pengalaman konsumen yang telah mengunjungi Pasar Lama Tangerang. Dalam penelitian ini dapat disimpulkan dengan menggunakan 120 data, yang terdiri dari 102 data digunakan sebagai data *training* dan 18 data digunakan sebagai data *testing*, dengan menggunakan *Confusion Matrix* menghasilkan akurasi sebesar 83% dengan presisi positif 100% dan presisi negative sebesar 40%. Dari 120 data yang menghasilkan 92 komentar positif dan 28 komentar negatif, dengan banyak komentar positif dapat menunjukkan bahwa Pasar Lama Tangerang layak untuk dikunjungi sebagai objek wisata kota Tangerang bagi masyarakat kota Tangerang dan di luar kota Tangerang sebagai konsumen.

Penelitian Onantya dkk. (2019) yang berjudul "Analisis Sentimen Pada Ulasan Aplikasi BCA Mobile Menggunakan BM25 dan *Improved K-Nearest Neighbors*". Data yang digunakan adalah data ulasan aplikasi BCA Mobile di toko aplikasi *playstore*.

Tujuan dari penelitian ini adalah untuk mengidentifikasi model algoritma perancangan algoritma BM25 dan *improved K-Nearest Neighbors* untuk mendapatkan *sentiment analysis* pada komentar aplikasi Mobile Banking BCA, serta melakukan validasi yang didapat melalui proses *precision, recall,* dan *f-measure (k-fold Cross validation).* Dalam penelitian ini dapat disimpulkan bahwa dengan metode tersebut, didapatkan rata – rata nilai yang cukup tinggi dengan nilai *f-meansure precision, recall* dan *accuracy* secara berturut – turut 0,939, 0,946, dan

0,934, dan 0,942. Besarnya nilai tersebut dipengaruhi oleh nilai *k-values* yang tepat dengan menyesuaikan data uji tiap kelasnya

Pada penelitian *sentiment analysis* terhadap tayangan televisi berdasarkan opini masyarakat pada media Twitter yang dilakukan oleh (Nurjanah et al., 2017). Tujuan penelitian ini adalah memberikan panduan dan pertimbangan untuk penonton tayangan televisi dalam menentukan tayangan yang digemari masyarakat pada umumnya. Dalam penelitian ini dapat disimpulkan bahwa dari 400 data yang dikumpulkan dan dianalisis dengan *KNN method* dan *scoring* jumlah *retweet* cukup mewakili untuk diaplikasikan dalam *analysis sentiment* masyarakat terhadap tayangan televisi. Hasil penelitian menunjukkan dengan nilai k optimal = 3 diperoleh tingkat akurasi, *precision*, dan *recall* masing – masing 80,83%, 72,28%, dan 100%.

Wibowo & Jumiati (2018) pernah melakukan penelitian Sentimen Analisis Masyarakat Pekalongan terhadap Pembangunan Jalan Tol Pemalang-Batang. Data yang digunakan berasal dari *group Facebook* Pekalongan Info. Tujuan dari penelitian ini untuk melihat sebuah opini masyarakat terutama warga Pekalongan yang ditunjukkan pada pembangunan jalan tol Pemalang-Batang. Dalam penelitian ini dapat disimpulkan bahwa dengan menggunakan *rapidminer* versi 7.6 dan menggunakan metode *Naïve Bayes* diperoleh nilai akurasi tertinggi mencapai 73,54% dengan nilai fold 3.

Pada penelitian ini diusulkan untuk membandingkan kinerja tiga buah algoritma klasifikasi dengan Teknik *Majority Vote*, Ketiga metode klasifikasi yang dimaksud adalah *SVM, KNN* dan *Naïve Bayes.* Diharapkan dapat menghasilkan hasil yang optimal.

2.3. Youtube API

Application Programming Interface adalah kumpulan perintah dan protokol untuk membantu dalam membangun atau mengembangkan *software* tertentu, dalam YouTube sendiri YouTube API dapat digunakan pengembang mengakses statistik video dan data saluran YouTube melalui dua jenis panggilan, REST dan XML-RPC.

Untuk menggunakan YouTube API, seorang pengembang harus memiliki Developer ID. Developer ID ini yang akan digunakan untuk mendapatkan akses statistik, komentar dalam sebuah video.

2.4. Text Mining

Text Mining adalah proses mengekstraksi informasi dan pengetahuan dari data yang tidak terstruktur. Data yang digunakan untuk *Text Mining* dibedakan menjadi dua yaitu data yang tidak terstruktur (dokumen Word, PDF, kutipan teks) dan data yang terstruktur (Balya, 2019).

Text Mining sering digunakan untuk mengkategorikan, dan untuk menganalisis sentiment yang berguna mengidentifikasi wawasan dan pola *trend* dalam *volume* besar dari data yang terstruktur. Banyak perusahaan menggunakan *Text Mining* untuk menganalisis pendapat, emosi dan *sentiment* dari masyarakat terhadap merek dan produk mereka.

2.5. Preprocessing Text mining

Preprocessing Text mining adalah proses menyeleksi data yang akan diproses sehingga menghasilkan data yang lebih baik dan sudah terstruktur dengan jelas (Balya, 2019).

Ada pun tahapan dalam Preprocessing yaitu *Cleaning*, *Case Folding*, *Tokenizing*, *Stopword Removal* dan Stemming (Maisarah, 2020). *Cleaning* data adalah proses menghapus *tweet special*, angka, tanda baca, simbol tidak penting dan *emoji*. *Case Folding* adalah proses mengubah huruf menjadi *lowercase*.

Sedangkan proses *Tokenizing* adalah proses untuk memecah sekumpulan

karakter yang dipisahkan oleh karakter *whitespace*. *Whitespace* adalah ruang kosong atau pemisah kata antara kalimat, contoh *whitespace,* seperti *enter,* tabulasi *spasi*. Untuk lebih jelasnya tahapan *Tokenizing* contoh tahapan dapat dilihat pada Tabel 2.1 di bawah ini.

Tabel 0.1 Contoh tahap *Tokenizing* Teks

Teks Input	Teks Output			
Great video thanks	Great	video	thanks	

Pada proses *Stopword Removal* adalah proses untuk membuang kata-kata yang kurang penting dari hasil *Tokenizing*, dengan menggunakan algoritma *stoplist* atau algoritma *wordlist*. Stoplist/stopword adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag-of-words*. Contoh *stopwords* adalah "*with*", "*and*", "then" dan seterusnya.

Preprocessing yang terakhir adalam Stemming. Stemming adalah proses mengubah kata menjadi kata dasar.

Tabel 0.2 Contoh tahap Stemming Teks.

Tokenizing				Stemming			
	hate	curve	screens		hate	curve	screens
	buying	samsung	again		buy	samsung	again

Dalam melakukan proses *Tokenizing*, *Stopword* Removal dan Ste*mming* dan phyton menggunakan *library NLTK. NLTK (The Natural Language Toolkit)* adalah rangkaian modul program, kumpulan data dan tutorial yang mendukung penelitian dan pengajaran dalam linguistik komputasi dan pemprosesan *natural language* (Loper & Bird, 2002).

2.6. Scraping

Scraping adalah Teknik untuk mengekstrak informasi dari sebuah website secara otomatis (Diaz & Rangkuti, 2020). *Scraping* berarti latihan membaca pada data teks dari layar terminal komputer. Dalam bentuknya saat ini, *scraping* merupakan bagian dari pemrograman yang menghubungkan antara program aplikasi dan antarmuka pengguna. Hal ini dirancang untuk berinteraksi dengan perangkat dan sistem antarmuka sehingga program tampilan yang berupa teks maupun gambar yang tidak mengandung fungsi dalam bentuk logika masih dapat dimanfaatkan sebagai data dan kemudian dapat diolah menjadi dataset.

2.7. Term Frequency – Invers Document Frequency (TF-IDF)

TF-IDF (Term Frequency – Invers Document Frequency) adalah pembobotan yang sering digunakan dalam pencarian informasi dan *text mining* (Alif, 2020). Bobot ini adalah ukuran statistik yang digunakan untuk mengevaluasi seberapa penting suatu kata dari sebuah dokumen dalam dataset. Pada penelitian (Shi & Li, 2011) terbukti bahwa hasil dari pembobotan menggunakan *TF-IDF* lebih efektif daripada frekuensi. Kombinasi dari nilai tf dan nilai idf dalam perhitungan bobot merupakan Nilai tf-idf dari sebuah kata. Persamaan TF-IDF dapat dilihat pada persamaan (2.1) sebagai berikut:

$$w_{dt} = tf_{dt} \times idf_{ft}$$
$$w_{dt} = tf_{ij} \times \log\left(\frac{D}{df_j}\right) \quad (2.1)$$

Keterangan

w_{ij} = Bobot term t_j terhadap dokumen d_i

tf_{ij} = Jumlah kemunculan term t_f dalam dokumen d_i

D = Jumlah semua dokumen yang ada dalam *database*

df_j = Jumlah dokumen yang mengandung term t_j

2.8. Support Vector Machine

Support Vector Machine (SVM) adalah satu dari metode klasifikasi untuk mengklasifikasi data yang akan dipisahkan dalam beberapa kelas (Zamahsyari & Nurwidyantoro, 2017). *Support Vector Machine* akan membuat garis pemisah di antara mereka untuk membedakan setiap kelas. Garis yang digunakan untuk memisahkan kelas disebut *separating*

hyperplan. Hyperplane hanyalah sebuah garis. Tapi, dalam dataset tiga dimensi, diperlukan pesawat untuk memisahkan data dan dalam dimensi lO24, sesuatu dengan dimensi lO23 diperlukan untuk memisahkan data, sesuatu disebut *hyperplane*. *Hyperplane* adalah batas keputusan, segala sesuatu di satu sisi milik satu kelas dan segala sesuatu di sisi lain milik kelas yang berbeda.

Tahapan pembelajaran dalam SVM adalah :

(1) Mencari Lagrance Multipliers(ai)

$$L(a) = \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i,j} a_i a_j y_i y_j X \frac{T}{i} x_j$$

$Dikarenakan\ (untuk\ setiap\ i = 1 \ldots\ldots n$

Keterangan

y_i = Kelas data latih (+1/-1).
y_j = Kelas data latih (+1/-1).
x_i = *Vector* bobot kalimat komentar.
x_j = *Vector* bobot kalimat komentar.

b) Mencari nilai Bobot(w)

$$w = \sum_{i=1}^{n} (\propto_i y_i x_i)$$

Keterangan

w = *Vector* bobot.
y_i = Kelas data latih(+1/-1).
x_i = *Vector* bobot kalimat komentar yang menjadi *vector* pendukung.

c) Mencari Nilai Bias(b)

$$b = \frac{1}{NSV} \sum_{i=1}^{NSV} (w.x_i - y_i)$$

Keterangan

NSV = Jumlah vector pendukung.
w = *Vector* bobot.
y_i = Kelas data latih(+1/-1).
x_i = *Vector* bobot kalimat komentar yang menjadi *vector* pendukung.

Proses pengklasifikasian (pengujian) dalam SVM menggunakan persamaan berikut.

$$f(t) = sign\left(\sum_{i=1, x \in SV}^{n} a_i y_i < t.x_i > + b \right) \quad (2.1)$$

Keterangan

T = *Vector* bobot data uji.
x_i = *Vector* pendukung.
b = Nilai bias.
y_i = Kelas atau label dari *vector* pendukung(+1/-1)

2.9. K-Nearest Neighbors

K-Nearest Neighbors (K-NN) adalah metode untuk mengklasifikasikan objek berdasarkan *data training* yang paling dekat dengan objek (Ernawati & Wati, 2018). Oleh karena itu, untuk menggunakan metode K-NN diperlukannya metrik untuk mengukur jarak antara titik *query* dan *sample*. Metrik yang paling populer mengukur jarak dikenal sebagai *Euclidean*. Persamaan *Euclidean* ditunjukkan pada persamaan (2.9) sebagai berikut:

$$D(x, p) = \sqrt{(x - p)^2} \quad (2.1)$$

Keterangan

x = poin *query* dari contoh kasus.
p = poin *query* dari contoh kasus.

Metode K-NN didasarkan pada asumsi intuitif bahwa objek yang jaraknya dekat berpotensi sama, maka dapat diasumsikan dengan jarak antar tetangga terdekat dengan titik yang dicari(K) ketika membuat prediksi. Tetangga yang terdekat akan memiliki lebih banyak *point*. Hal ini dapat dilakukan jika menggunakan bobot(W) untuk masing – masing tetangga terdekat. Bobot itu akan ditentukan oleh kedekatan relatif masing – masing tetangga dengan K.

$$W(x_0, x_i) y_i = \frac{exp(-D(xi, pi))}{\sum_{i=1}^{k} exp(-D(xi, pi))} \quad (2.1)$$

Keterangan

$D(xi, pi)$ = jarak antara point x yang ada di kasus dengan contoh *sample pi*
x = poin *query* dari contoh kasus.
p = poin *query* dari contoh kasus.

Untuk menghitung bobot dapat dilakukan dengan cara berikut:

$$\sum_{i=1}^{k} W(x_0, x_i) = 1$$

Sehingga dalam masalah klasifikasi nilai maximum y adalah banyaknya kelas variable.

$$\max\left(y = \sum_{i=1}^{k} W(x_0, x_i)\, yi\right) \qquad (2.1)$$

Validasi yang digunakan pada penelitian ini adalah k-fold *Cross validation*. K-fold *Cross validation* membagi data akan dibagi menjadi k bagian (Wibowo & Jumiati, 2018).

2.10. Naïve Bayes

Naïve Bayes adalah salah satu algoritma klasifikasi *data mining*. Klasifikasi adalah proses untuk mengelompokkan data berdasarkan keterkaitan data terhadap data sample (Balya, 2019). *Naïve Bayes* menggunakan perhitungan probabilitas yang menggunakan konsep pendekatan Bayesian. *Naïve Bayes* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam *database* dengan data yang besar (Zulfikar & Lukman, 2016).

$$P(cj|wi) = \frac{P(wi|cj)P(cj)}{P(wi)}$$

Keterangan

$P(cj\|wi)$	=	Peluang kategori j, Ketika terdapat kemunculan kata i
$P(wi\|cj)$	=	Peluang kata i masuk ke dalam kategori j
$P(cj)$	=	Peluang kemunculan kategori j
$P(wi)$	=	Peluang kemunculan kata

2.11. Majority vote

Ketika terdapat tiga metode klasifikasi yang berbeda, hl (X), h2 (X), dan h3 (X). Ketika metode tersebut dapat digabungkan sehingga dapat menghasilkan metode klasifikasi yang lebih unggul daripada aturan individual dalam keadaan tertentu. Cara umum untuk membandingkan aturan ini adalah:

$$C(X) = mode\{h_1(X), h_2(X), h_3(X)\} \qquad (2.2)$$

Dengan kata lain, setiap nilai X adalah jumlah klasifikasi dari metode klasifikasi. Penggabungan klasifikasi ini dikenal sebagai *Majority Vote Classifiers* atau *Majority vote Learners.*

2.12. Python

Python adalah bahasa pemrograman yang bersifat open source. *Python* telah dioptimalkan untuk software quality, developer productivity origram *portability* dan *component integration* (Harismawan et al., 2018). Python Bahasa yang mudah dipelajari karena memiliki sintaks yang jelas, memiliki banyak modul atau library yang sudah siap digunakan, dan memiliki tingkat keakuratan data yang tinggi dan efisien (Dedi Ary Prasetya, 2012).

Dalam perbandingan bahasa pemrograman antara Python, Php, dan Perl, Python memiliki keunggulan dalam penggunaan *memory* dan *Cpu* paling sedikit dibandingkan dengan Bahasa pemrograman lainnya (Harismawan et al., 2018).

2.13. Jupyter Notebook

Notebook dirancang untuk mendukung alur kerja komputasi ilmiah, dari eksplorasi interaktif untuk menerbitkan catatan komputasi terperinci. Kode dalam file Notebook diatur ke dalam sel, yang setiap potongnya dapat dimodifikasi dan dijalankan secara individual. Keluaran dari setiap sel muncul tepat di bawahnya dan disimpan sebagai bagian dari dokumen. Ini adalah evolusi dari shell interaktif atau REPL (read-evalu-print loop) yang telah lama menjadi dasar pemrograman interaktif.

Antarmuka Notebook pertama kali menjadi populer di kalangan matematikawan. Jupyter memiliki tujuan untuk menghadirkan Notebook ke pengguna yang lebih luas. Jupyter adalah open source proyek, yang dapat bekerja dengan kode dalam berbagai bahasa pemrograman (Kluyver dkk., 2016).

Jupyter *Notebook* dapat diakses melalui *web browser*. Dengan memiliki antarmuka yang sama antara berjalan secara offline seperti aplikasi *desktop*, atau berjalan di

server membuatnya menjadi mudah digunakan. *File Notebook* yang digunakan adalah format JSON yang terdokumentasi dan sederhana, dengan ekstensi '.ipynb'. Sangat mudah untuk menulis perangkat lunak lain yang mengakses dan memanipulasi suatu file.

2.14. Cross validation

Cross validation adalah metode untuk mengevaluasi dan membandingkan pembelajaran dari algoritma (learning algorithms) dengan membagi data menjadi dua bagian, satu bagian digunakan untuk *training* dan bagian lainnya digunakan sebagai *testing* (Wibowo & Jumiati, 2018).

2.15. Teknik Imbalance

Data *imbalance* adalah kondisi dimana tidak seimbangnya kategori klasifikasi (Julia, 2018). Ia juga mengatakan bahwa teknik *sampling* adalah salah satu metode yang paling populer untuk mengatasi ketidakseimbangan kelas. Pengujian dataset yang tidak seimbang memiliki *misclassification cost* (nilai *instance* yang terklasifikasi) yang dimiliki kelas minoritas lebih tinggi dari pada kelas mayoritas.

Salah satu teknik *sampling* adalah *undersampling.* Teknik *undersampling* adalah teknik dimana metode yang digunakan berupa *re-sampling* eliminasi dari kelas mayoritas secara acak sampai jumlahnya sebanyak kelas lainnya atau kelas minoritas (Julia, 2018).

2.16. Evaluasi Sistem

Evaluasi Sistem Evaluasi sistem dilakukan dengan cara menghitung tingkat keakuratan suatu metode untuk menganalisis hasil komentar dari Youtube (Kurniawan & Susanto, 2019). Akurasi adalah fraksi prediksi model yang benar. Rumus akurasi, sebagai berikut:

$$Akurasi = \frac{Jumlah\ prediksi\ benar}{Jumlah\ data} \quad (2.1)$$

3. METODOLOGI

3.1. Alur Penelitian

Metodologi penelitian merupakan tahapan-tahapan yang dilakukan selama penelitian berlangsung agar pelaksanaan dapat tersusun dengan baik dan sistematis agar dapat mencapai tujuan seperti yang diharapkan. Pada penelitian ini diperlukan langkah-langkah kegiatan penelitian untuk mendapatkan hasil yang maksimal seperti yang ditunjukkan pada Gambar 3.1.

Gambar 1 Alur Penelitian

3.2. Identifikasi Masalah

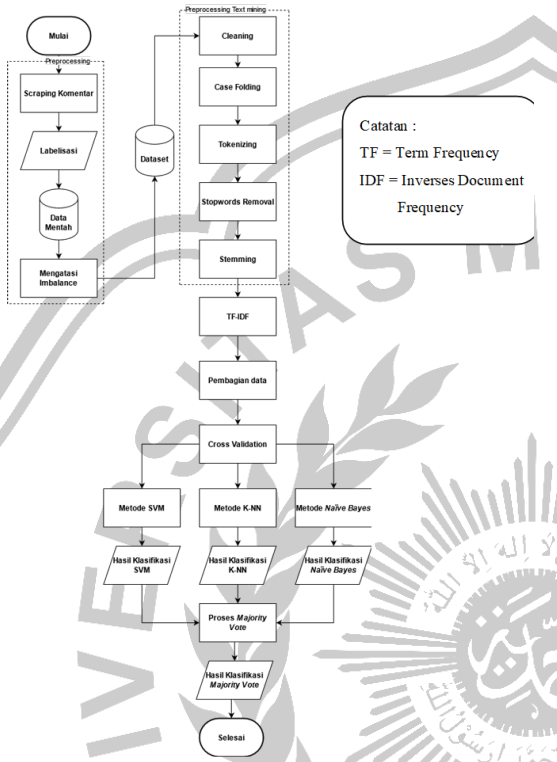
Tahapan awal setelah menentukan topik penelitian yang akan diambil adalah mengidentifikasi permasalah yang akan dipelajari. Tahapan Identifikasi Masalah dimaksud sebagai penegasan batas-batas permasalahan, sehingga cakupan penelitian tidak keluar dari tujuan. Dalam penelitian ini, proses identifikasi masalah tentang komentar penonton video *review* Samsung Galaxy S20 di YouTube, yang berupa singkatan, Bahasa asing, dan berbagai masalah dalam melakukan *sentiment analysis* pada YouTube. Untuk itu diperlukan Algoritma yang mampu menemukan pesan yang terkandung dalam komentar penonton di YouTube dan sekaligus dapat menyeleksi kata yang tidak diperlukan dan mengubahnya menjadi data yang dapat dijadikan fitur.

3.3. Studi Literatur

Pada tahap ini akan dilakukan dengan cara mengumpulkan bahan referensi yaitu berupa buku, artikel, paper, jurnal, dan makalah yang berkaitan dengan metode TF-IDF, Sentiment Analysis, Algoritma Support Vector Machine, Algoritma K-Nearest Neighbors, Algoritma Naïve Bayes dan Metode Majority Vote.

3.4. Desain Sistem

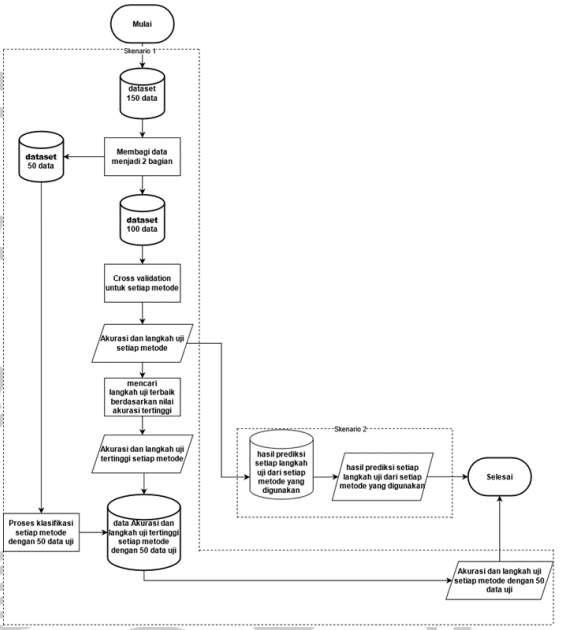
Proses perancangan desain sistem dapat dilihat pada gambar 3.2 berikut.



Gambar 2 Alur Sistem

3.5. Skenario pengujian

Skenario pengujian adalah skenario yang akan dilakukan untuk menjawab rumusan masalah pada bab 1 dengan menggunakan 150 data. Pada penelitian ini terdapat 2 skenario yaitu:



Gambar 3 Skenario pengujian

1. Skenario perbandingan akurasi yang dihasilkan *Majority vote* dibandingkan metode *SVM, KNN* dan *Naïve Bayes.*

   Pada skenario ini akurasi *Majority vote*, *SVM,KNN* dan *Naïve Bayes* yang dibandingkan berasal dari akurasi dari langkah uji terbaik masing - masing metode yang didapatkan melalui metode *cross validation,* sebelum menggunakan metode *cross validation* data akan di pisah menjadi 2 data yaitu 100 data yang akan digunakan dalam metode *cross validation* untuk mencari Langkah iji terbaik dan 50 data digunakan sebagai data uji, tujuan memisahkan data adalah untuk mengetahui apakah langkah uji terbaik yang telah didapatkan melalui proses cross validation tetap mendapatkan akurasi yang tetap mendapatkan baik atau tidak.

2. Skenario mencari metode yang terpilih

dalam metode *Majority vote.*

Pada skenario ini pencarian metode didapatkan dari menghitung metode yang terpilih oleh *Majority vote* dari seluruh pengujian skenario 1.

4. HASIL DAN PEMBAHASAN

Data yang didapatkan sebanyak 240 yang terdiri dari komentar 50 komentar positif, 68 komentar negatif, 122 komentar netral, yang disembangkan sehingga didapatkan data sebanyak 150 data yang terdiri dari komentar 50 komentar positif, 50 komentar negatif, 50 komentar netral. 150 data tersebut dilakukan berbagai proses seperti *cleaning, case folding, tekonizing, stopword removal, stemming*. Selanjutnya proses ekstraksi fitur menggunakan TFIDF agar dapat dihitung menggunakan metode, setelah proses tersebut dibagi menjadi 2 bagian 1. Bagian berjumlah 100 data, bagian 2 berjumlah 50 data, 100 akan dilakukan pembagian data atau partisi data dengan menggunakan K Fold Cross Validation untuk mendapatkan Langkah uji terbaik. Setelah dilakukan pembagian data maka masuk kedalam tahap klasifikasi masing masing metode ditampilkan Gambar 4

Langkah uji	*Support Vector Machine*	*K-Nearest Neighbors*	*Naïve Bayes*	*Majority Vote*
2-fold langkah uji 1	0.24	0.54	0.44	0.42
2-fold langkah uji 2	0.36	0.32	0.4	0.34
4-fold langkah uji 1	0.36	0.56	0.48	0.52
4-fold langkah uji 2	0.16	0.36	0.28	0.24
4-fold langkah uji 3	0.4	0.28	0.36	0.36
4-fold langkah uji 4	0.28	0.28	0.38	0.28
5-fold langkah uji 1	0.5	0.5	0.45	0.45
5-fold langkah uji 2	0.45	0.6	0.6	0.55
5-fold langkah uji 3	0.4	0.6	0.55	0.65
5-fold langkah uji 4	0.45	0.35	0.3	0.3
5-fold langkah uji 5	0.3	0.2	0.4	0.3
10-fold langkah uji 1	0.5	0.3	0.5	0.5
10-fold langkah uji 2	0.3	0.3	0.5	0.4
10-fold langkah uji 3	0.5	0.7	0.6	0.6
10-fold langkah uji 4	0.5	0.8	0.5	0.8
10-fold langkah uji 5	0.3	0.2	0.4	0.2
10-fold langkah uji 6	0.5	0.7	0.4	0.5
10-fold langkah uji 7	0.6	0.6	0.4	0.6
10-fold langkah uji 8	0.1	0	0.3	0

Gambar 4 Akurasi seluruh Langkah uji

50 data digunakan untuk menguji Langkah uji terbaik yang didapatkan masing-masing metode. Setelah dilakukan perhitungan klasifikasi masing-masing metode berdasarkan Langkah uji terbaik yang diuji dengan 50 data uji ditampilkan Gambar 5

Metode	Akurasi
Support Vector Machine 10 fold langkah uji 7	0.36
K-Nearest Neighbors 10 fold langkah uji 4	0.48
Naïve Bayes 5 fold langkah uji 2	0.52
Majority vote 10 fold langkah uji 4	0.48

5. KESIMPULAN DAN SARAN

5.1. KESIMPULAN

Kesimpulan yang dapat penulis berikan selama melakukan penelitian dan pembuatan sistem mengenai sentiment analysis komentar YouTube samsung s20 menggunakan metode Majority vote yaitu:

a) Tingkat akurasi Majority vote berdasarkan langkah uji 10-fold langkah uji 4 menggunakan 100 data mendapatkan nilai akurasi 0.8 dan berdasarkan langkah uji 10-fold langkah uji 4 dengan 50 data uji mendapatkan nilai akurasi 0.48.
b) Tingkat akurasi Majority vote lebih baik dari Support Vector Machine, tetapi tidak lebih baik dari Naïve Bayes dan memiliki tingkat akurasi yang sama dengan K-Nearest Neighbors.
c) Metode yang paling sering terpilih dalam metode Majority vote pada semua langkah uji yang telah dilakukan adalah Metode K-Nearest Neighbors yaitu terpilih sebanyak 319 kali.

5.2. SARAN

Saran yang dapat penulis sampaikan setelah melakukan penelitian ini untuk meningkatkan pengembangan dan kualitas pada penelitian yang akan datang yaitu:

a) Menggunakan metode yang lainnya

untuk dibandingkan dalam Majority Vote.

b) Menggunakan metode pembagian data selain test split dan Cross validation yaitu Leave-one-out cross-validation (LOOCV) dan Stratified K-fold cross-validation.

c) Dapat melakukan klasifikasi sentimen analisis pada emoticon.

DAFTAR PUSTAKA

Alif, F. Z. (2020). Ekstraksi Fitur untuk Pemilihan Topik Spesifik Review Film dalam Menghasilkan Aspect-Based Sentiment Analysis.pdf. Institusi USU, Universitas Sumatera Utara.

Balya. (2019). Analisis Sentimen Pengguna Youtube di Indonesia pada Review Smartphone Menggunakan Naïve Bayes.

Dedi Ary Prasetya, I. N. (2012). Deteksi wajah metode viola jones pada opencv menggunakan pemrograman python. Simposium Nasional RAPI XI FT UMS, 18–23.

Diaz, M., & Rangkuti, E. (2020). Analisis topik komentar video beberapa akun youtube e-commerce indonesia menggunakan metode latent dirichlet allocation.

Ernawati, S., & Wati, R. (2018). Penerapan Algoritma K-Nearest Neighbors Pada Analisis Sentimen Review Agen Travel. Jurnal Khatulistiwa Informatika, VI(1), 64–69. https://ejournal.bsi.ac.id/ejurnal/index.php/khatulistiwa/article/view/3802/2626.

Harismawan, A. F., Kharisma, A. P., & Afirianto, T. (2018). Analisis Perbandingan Performa Web Service Menggunakan Bahasa Pemrograman Python , PHP , dan Perl pada Client Berbasis Android. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya, 2(January), 237–245. http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/781

Hustinawaty, Dwiputra, R. A. A., & Rumambi, T. (2019). Public Sentiment Analysis of Pasar Lama Tangerang Using K-Nearest Neighbor Method and Programming Language R. Jurnal Ilmiah Informatika Komputer, 24(2), 129–133. https://doi.org/10.35760/ik.2019.v24i2.2367

Julia, W. (2018). Klasifikasi Pembiayaan Warung Mikro Menggunakan Metode Random Forest Dengan Teknik Sampling Kelas Imbalanced. Высшей Нервной Деятельности, 2, 227–249.

KBBI. (2016). Kamus Besar Bahasa Indonesia (KBBI). In Kementerian Pendidikan dan Budaya.

Kurniawan, I., & Susanto, A. (2019). Implementasi Metode K-Means dan Naïve Bayes Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019. Eksplora Informatika, 9(1), 1–10. https://doi.org/10.30864/eksplora.v9i1.237

Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit. July, 69–72. https://doi.org/10.3115/1225403.1225421

Maisarah, M. A. (2020). Sistem Analisis Sentimen pada Fanpage Facebook Kandidat Presiden 2019-2024.

Næs, T., Lengard, V., Bølling Johansen, S., &

Hersleth, M. (2010). Alternative methods for combining design variables and consumer preference with information about attitudes and demographics in conjoint analysis. Food Quality and Preference, 21(4), 368–378. https://doi.org/10.1016/j.foodqual.2009.09.004

Nurjanah, W. E., Perdana, R. S., & Fauzi, M. A. (2017). Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya, 1(12), 1750–1757. https://doi.org/10.1074/jbc.M209498200

Onantya, I. D., Indriati, & Adikara, P. P. (2019). Analisis Sentimen Pada Ulasan Aplikasi BCA Mobile Menggunakan BM25 Dan Improved K-Nearest Neighbor. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 3(3), 2575–2580.

Rintyarna, B. S. (2016). Sentiment Analysis pada Data Twitter dengan Pendekatan Naïve Bayes Multinomial. 1–6.

Rosdiansyah, D. (2014). Analisis Sentimen Twitter Menggunakan Metode K-Nearest Neighbor dan Pendekatan Lexicone. Tugas Akhir Jurusan Teknik Informatika, 1–15.

Shi, H. X., & Li, X. J. (2011). A sentiment analysis model for hotel reviews based on supervised learning. Proceedings - International Conference on Machine Learning and Cybernetics, 3, 950–954. https://doi.org/10.1109/ICMLC.2011.6016866

Sipayung, E. M., Maharani, H., & Zefanya, I. (2016). Perancangan Sistem Analisis Sentimen Komentar Pelanggan Menggunakan Metode Naive Bayes Classifier. Jurnal Sistem Informasi, 8(1), 958–965.

Tanesab, F. I., Sembiring, I., & Purnomo, H. D. (2017). Sentiment Analysis Model Based On Youtube Comment Using Support Vector Machine. International Journal of Computer Science and Software Engineering (IJCSSE), 6(8), 180–185. http://ijcsse.org/published/volume6/issue8/p2-V6I8.pdf

Wibowo, A. P., & Jumiati, E. (2018). Sentiment Analysis Masyarakat Pekalongan Terhadap Pembangunan Jalan Tol Pemalang-Batang Di Media Sosial. IC-Tech, XIII(0285), 42–48.

Zamahsyari, & Nurwidyantoro, A. (2017). Sentiment analysis of economic news in Bahasa Indonesia using majority vote classifier. Proceedings of 2016 International Conference on Data and Software Engineering, ICoDSE 2016. https://doi.org/10.1109/ICODSE.2016.7936123

Zulfikar, W. B., & Lukman, N. (2016). Perbandingan Naive Bayes Classifier Dengan Nearest Neighbor Untuk Identifikasi Penyakit Mata. Jurnal Online Informatika, 1(2), 82–86. https://doi.org/10.15575/join.v1i2.33