

SKRIPSI

**IMPLEMENTASI SENTIMEN ANALISIS PADA ULASAN
PENGGUNA UNTUK REKOMENDASI E-WALLET**

Disusun dan Diajukan Oleh:

AINUN ANNISA KAHAR

D121181024



DEPARTEMEN TEKNIK INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS HASANUDDIN

2022

LEMBAR PENGESAHAN SKRIPSI
IMPLEMENTASI SENTIMEN ANALISIS PADA ULASAN PENGGUNA
UNTUK REKOMENDASI E-WALLET

Disusun dan diajukan oleh

AINUN ANNISA KAHAR

D121181024

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Studi Program Sarjana Program Studi Teknik Informatika Fakultas Teknik Universitas Hasanuddin pada tanggal 19 Oktober 2022 dan dinyatakan telah memenuhi syarat kelulusan.

Menyetujui,

Pembimbing Utama,



Dr. Amri Ahmad Ilham, S.T., MIT.
Nip. 197310101998021001

Pembimbing Pendamping,



Anugrayani Bustamin, S.T., M.T.
Nip. 199012012018074001

Ketua Program Studi,



Dr. Indrabayu, S.T., M.T.
Nip. 19750716 200212 1 004

PERNYATAAN KEASLIAN

Yang bertanda tangan dibawah ini:

Nama : Ainun Annisa Kahar
Nim : D121181024
Program Studi : Teknik Informatika
Jenjang : S1

Menyatakan dengan ini karya tulisan saya berjudul:

IMPLEMENTASI SENTIMEN ANALISIS PADA ULASAN PENGGUNA UNTUK REKOMENDASI E-WALLET

Adalah karya tulisan saya sendiri dan bukan merupakan pengambilan alihan tulisan orang lain bahwa skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri.

Apabila di kemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan skripsi ini hasil karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Gowa, 24 Oktober 2022

Yang Menyatakan,



Ainun Annisa Kahar

KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada Allah SWT atas rahmat dan berkat-Nya, sehingga dapat menyelesaikan tugas akhir skripsi yang berjudul **“Implementasi Sentimen Analisis Pada Ulasan Pengguna Untuk Rekomendasi E-Wallet”** sebagai salah satu syarat dalam menyelesaikan jenjang Strata-1 di Departemen Teknik Informatika, Fakultas Teknik, Universitas Hasanuddin.

Penulis menyadari banyak kesulitan dan kendala yang dihadapi saat penyusunan tugas akhir ini. Dalam prosesnya, penulis memperoleh banyak bantuan, dukungan, dan bimbingan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan terima kasih kepada:

1. Allah SWT melalui berkat dan rahmat-Nya sehingga penulis dapat menyelesaikan tugas akhir ini.
2. Kedua orang tua penulis, Ayah Kahar Udin Kagas dan Ibu Misnawati Rahman yang selalu menyertai penulis dalam doanya serta mendukung, membantu, memberi semangat serta kasih sayang dalam perjalanan penulis menyelesaikan tugas akhir ini.
3. Bapak Dr. Amil Ahmad Ilham, S.T., M.IT. selaku pembimbing I dan Ibu Anugrayani Bustamin, S.T., M.T. selaku pembimbing II, yang senantiasa menyediakan waktu, tenaga, pikiran, dan perhatian yang luar biasa dalam mengarahkan penulis untuk menyelesaikan tugas akhir.
4. Segenap Dosen dan Staff Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin yang telah banyak membantu penulis selama masa perkuliahan.

5. Teman-teman Teknik Informatika Angkatan 2018 (Synchronous) selaku rekan yang telah memberi bantuan, dukungan dan semangat selama masa perkuliahan dan penyusunan tugas akhir ini.
6. Teman-teman sejak maba Tamara Auliani, Ayu Ade, Dea Nurhikma, Rahmadani, Iffat Ainiyyah, Rofifah Nurul, atas dukungan dan bantuannya selama pengerjaan skripsi.
7. Serta berbagai pihak atas segala dukungan dan bantuannya yang tidak dapat penulis tuliskan satu persatu.

Penulis berharap semoga Allah SWT berkenan membalas segala kebaikan yang telah diterima oleh penulis dari berbagai pihak yang telah membantu mempermudah penulis dalam mengerjakan tugas akhir ini. Penulis menyadari bahwa tugas akhir ini masih jauh dari kata sempurna, oleh karena itu penulis mengharapkan segala bentuk saran serta masukan yang membangun dari berbagai pihak. Semoga tugas akhir ini dapat memberikan pengetahuan dan manfaat bagi penulis dan pembaca.

Makassar, September 2022

Penulis,

Ainun Annisa Kahar

ABSTRAK

Peningkatan tren dan penggunaan *e-wallet* sejak pandemi meningkat secara pesat sebagai metode pembayaran akibat dibatasinya aktivitas masyarakat. Calon pengguna menghabiskan banyak waktu dalam memilih karena sangat bergantung pada rekomendasi dari mulut ke mulut sebelum menggunakan suatu produk. Ulasan yang tersedia secara *online* terlalu banyak untuk dianalisis secara manual, sehingga dilakukan analisis sentimen untuk mengklasifikasikan ulasan tersebut. Dengan tujuan tersebut, penelitian ini membangun tiga model klasifikasi machine learning yaitu Multinomial Naïve Bayes, Jurafsky Naïve Bayes, dan Support Vector Machine untuk mengklasifikasikan sentiment ulasan pengguna *e-wallet*. Evaluasi model akan dilakukan untuk membandingkan hasil sentimen dari komposisi data latih dan data uji yang sama. Model Ensemble Hard Voting dengan Modifikasi dibangun untuk mengkombinasikan hasil prediksi dari ketiga model. Hasil pengujian menunjukkan bahwa model Multinomial Naïve Bayes memiliki tingkat akurasi tertinggi yang mencapai 80,3%, diikuti oleh model Ensemble Hard Voting dengan modifikasi mencapai 80,1%, Jurafsky Naïve Bayes mencapai 79,4% dan Support Vector Machine mencapai 79,2%. Namun demikian, model Ensemble Hard Voting dengan Modifikasi menunjukkan belum mampu meningkatkan performa akurasi yang lebih tinggi dibandingkan dengan model individual Multinomial Naïve Bayes. Dari hasil tersebut maka diputuskan menggunakan model Multinomial Naïve Bayes yang memiliki tingkat akurasi tertinggi dalam mengklasifikasikan sentimen ulasan pengguna untuk rekomendasi *e-wallet*.

Kata kunci: analisis sentimen, naïve bayes, support vector machine, ensemble learning

DAFTAR ISI

HALAMAN PENGESAHAN	ii
HALAMAN PERNYATAAN KEASLIAN	iii
KATA PENGANTAR	iv
ABSTRAK	vi
DAFTAR ISI	vii
DAFTAR GAMBAR	ix
DAFTAR TABEL	x
DAFTAR LAMPIRAN	xi
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Tujuan Penelitian	3
1.4 Manfaat Penelitian	3
1.5 Batasan Masalah	4
1.6 Sistematika Penulisan	4
BAB II TINJAUAN PUSTAKA	5
2.1 E-Wallet	5
2.2 Analisis Sentimen	5
2.3 Bag of Words	6
2.4 Algoritma Naïve Bayes	7
2.5 Multinomial Naïve Bayes	8
2.6 Optimasi Naïve Bayes Untuk Analisis Sentimen	9
2.7 Support Vector Machine	10
2.8 Ensemble Learning	13
2.9 Confusion Matrix	13
BAB III METODOLOGI PENELITIAN	17
3.1 Tahapan Penelitian	17

3.2	Rancangan Sistem	18
3.3	Pengumpulan Data.....	20
3.4	Preprocessing Data	21
3.5	Analisis Sentimen Dengan Naïve Bayes	23
3.6	Analisis Sentimen Dengan Optimasi Naïve Bayes oleh Jurafsky dan Martin	27
3.7	Analisis Sentimen Dengan Support Vector Machine	33
3.8	Hard Voting Dengan Modifikasi Berdasarkan Perbandingan Performa Model	39
BAB IV HASIL DAN PEMBAHASAN.....		43
4.1	Pengumpulan Data.....	43
4.2	Preprocessing Data	44
4.3	Analisis Sentimen Dengan Multinomial Naïve Bayes	47
4.4	Analisis Sentimen Dengan Optimasi Naïve Bayes Jurafsky.....	48
4.5	Analisis Sentimen Dengan Support Vector Machine	50
4.6	Hasil Hard Voting Dengan Modifikasi.....	51
4.7	Perbandingan Performa Model.....	53
4.8	Prediksi Sentimen Ulasan Untuk Rekomendasi E-Wallet.....	54
BAB V KESIMPULAN DAN SARAN		59
5.1	Kesimpulan.....	59
5.2	Saran	60
DAFTAR PUSTAKA		61
LAMPIRAN.....		64

DAFTAR GAMBAR

Gambar 2.7.1 Hyperplane antara kelas positif dan negatif (Nugroho et al., 2005)	11
Gambar 2.8.1 Ensembel Hard Voting (Géron, 2019)	13
Gambar 2.9.1 Confusion matrix: (a) <i>Binary classification</i> (b) <i>Multiclass classification</i> (Markoulidakis et al., 2021).....	14
Gambar 3.1.1 Tahapan penelitian	17
Gambar 3.2.1 Rancangan Sistem	19
Gambar 3.5.1 Flowchart algoritma Multinomial Naive Bayes	23
Gambar 3.5.2 Algoritma Multinomial Naïve Bayes (Shiri, 2004).....	24
Gambar 3.6.1 Algoritma optimasi Naïve Bayes dengan negasi “_NOT” (Wahyudiarto, 2022)	28
Gambar 3.7.1 Flowchart analisis sentimen dengan SVM	34
Gambar 3.7.2 Algoritma Pegasos SVM (Johansson, 2016).....	35
Gambar 3.8.1 Algoritma Hard Voting dengan modifikasi.....	41
Gambar 4.1.1 Hasil scraping ulasan menggunakan Google Play Scraper	43
Gambar 4.8.1 Sentimen ulasan aplikasi Dana.....	54
Gambar 4.8.2 Sentimen ulasan aplikasi iSaku	55
Gambar 4.8.3 Sentimen ulasan aplikasi Jenius	55
Gambar 4.8.4 Sentimen ulasan aplikasi LinkAja.....	56
Gambar 4.8.5 Sentimen ulasan aplikasi OVO	57

DAFTAR TABEL

Tabel 3.5.1 Sampel data untuk klasifikasi sentimen menggunakan Naive Bayes	25
Tabel 3.8.1 Contoh cara kerja Ensemble Hard Voting pada dua kelas.....	40
Tabel 3.8.2 Contoh Ensemble Hard Voting pada tiga kelas	40
Tabel 4.1.1 Sampel ulasan pengguna aplikasi OVO.....	44
Tabel 4.2.1 Hasil <i>preprocessing</i> data pada ulasan	45
Tabel 4.2.2 Rincian dataset	46
Tabel 4.3.1 Confusion Matrix Multinomial NB.....	47
Tabel 4.3.2 Evaluasi performa Multinomial NB.....	47
Tabel 4.4.1 Confusion Matrix Jurafsky NB	48
Tabel 4.4.2 Evaluasi performa Jurafsky NB	49
Tabel 4.5.1 Confusion matrix SVM.....	50
Tabel 4.5.2 Evaluasi performa SVM.....	50
Tabel 4.6.1 Confusion Matrix Hard Voting dengan modifikasi	52
Tabel 4.6.2 Evaluasi performa Hard Voting dengan modifikasi	52
Tabel 4.7.1 Perbandingan performa model	53

DAFTAR LAMPIRAN

Lampiran 1. Hasil prediksi model Multinomial NB, Jurafsky NB, SVM, dan Hard Voting dengan Modifikasi.....	64
Lampiran 2 Hasil prediksi Hard Voting dengan Modifikasi.....	64
Lampiran 3. Preprocess Data.....	65

BAB I

PENDAHULUAN

1.1 Latar Belakang

Peningkatan tren dan penggunaan *e-wallet* sejak pandemi meningkat secara pesat sebagai metode pembayaran akibat dibatasinya aktivitas masyarakat. Peningkatan ini diiringi oleh berbagai ulasan pengguna secara *online* terutama di ulasan aplikasi *e-wallet* pada Google Play store. OVO, Dana, Payfazz, Linkaja dan Isaku adalah beberapa *e-wallet* yang memiliki jumlah unduhan terbanyak di Google Play Store (Basryah et al., 2021).

Bank Indonesia menyatakan bahwa pada tahun 2020 terdapat 47 perusahaan *e-wallet* yang telah memiliki izin resmi (Kristiyanti et al., 2020). Perusahaan bersaing dalam menawarkan promosi serta fitur menarik yang mereka miliki kepada pengguna karena semakin banyak produk *e-wallet* yang berkembang. Masyarakat termasuk pengguna dan pemilik toko menghabiskan banyak waktu untuk memilih *e-wallet* sebagai metode pembayaran terbaik mereka. Karena banyaknya pertimbangan dalam melihat rating, ulasan, reputasi, keamanan, serta fitur tambahan lainnya. Rekomendasi dapat menjadi solusi yang membantu masyarakat dalam memilih.

Menurut Sung, sebanyak 50% pengguna internet bergantung pada rekomendasi *word-of-mouth* sebelum menggunakan suatu produk karena ulasan yang ditulis pengguna secara *online* lebih relevan terhadap informasi pengalaman mereka yang telah menggunakan produk tersebut (Utomo et al., 2021). Namun ulasan yang tersedia secara *online* terlalu banyak untuk dianalisis secara manual.

Analisis sentimen dapat digunakan untuk mengklasifikasikan ulasan tersebut. Analisis terhadap ulasan dapat dilakukan dengan melihat *rating* produk, namun *rating* saja belum cukup untuk mewakili maksud dari keseluruhan ulasan oleh pelanggan (Muktafin et al., 2020). Dalam rekomendasi, ulasan pengguna adalah salah satu alat yang sangat penting. Pengguna biasanya diberikan rekomendasi berdasarkan riwayat dan profil pengguna, namun saat ini rekomendasi mulai lebih banyak dieksplorasi sehingga membuat rekomendasi berdasarkan sentimen analisis (Khanvilkar and Vora, 2019). Dalam sentimen analisis ada beberapa klasifikasi algoritma Machine Learning yang digunakan yaitu seperti algoritma Multinomial Naive Bayes, Logistic Regression, Support Vector Machine Classifier, Decision Tree dan Random Forest (Khanvilkar & Vora, 2019b).

Penelitian yang telah dilakukan Kristiyanti menganalisis sentimen ulasan pengguna OVO dan Dana menunjukkan algoritma NB unggul pada sentimen negatif sedangkan SVM unggul pada sentimen positif (Kristiyanti et al., 2020). Himawan (Himawan et al., 2018) mengekstrak dan menganalisis opini masyarakat menggunakan pendekatan Lexicon Based dan SVM sehingga dapat menghasilkan rekomendasi oleh-oleh favorit Yogyakarta. Wahyudiarto pada tahun 2022 mengembangkan sebuah sistem rekomendasi produk e-commerce dengan mengusulkan model baru yang meningkatkan performa pada algoritma Naive Bayes yang konvensional (Wahyudiarto, 2022). Secara keseluruhan hasil yang diberikan mencapai 91% dengan nilai F1-score 57%. Namun model yang diusulkan hanya unggul pada sentimen negatif dan tidak demikian pada sentimen netral dan positif. Pada tugas akhir ini juga akan mengembangkan sistem yang

serupa. Perbedaannya adalah algoritma klasifikasi yang akan digunakan yaitu Multinomial Naïve Bayes, Jurafsky Naïve Bayes, dan Support Vector Machines untuk mengklasifikasikan sentimen. Model Ensemble Hard Voting dengan modifikasi juga digunakan untuk mengkombinasikan hasil prediksi dari ketiga model klasifikasi tersebut dan bertujuan untuk mencapai kinerja yang lebih andal dan akurat dibandingkan dengan model sebelumnya. Kemudian akan dipilih model terbaik untuk menghasilkan output sentimen yang kemudian dapat digunakan untuk memberikan rekomendasi *e-wallet*.

1.2 Rumusan Masalah

1. Bagaimana membangun model analisis sentimen terhadap ulasan pengguna *e-wallet*?
2. Bagaimana memberikan rekomendasi *e-wallet* berdasarkan sentimennya?

1.3 Tujuan Penelitian

1. Untuk membangun model analisis sentimen terhadap ulasan pengguna *e-wallet*.
2. Untuk memberikan rekomendasi *e-wallet* berdasarkan sentimennya.

1.4 Manfaat Penelitian

Penelitian ini diharapkan dapat membantu:

1. Masyarakat, terutama pemilik toko atau *merchant* dalam memilih metode pembayaran menggunakan *e-wallet*

2. Pengguna *e-wallet*, dengan hadirnya inovasi dan kreatifitas dalam memanfaatkan digitalisasi dalam bidang finansial bisnis
3. Perusahaan *e-wallet*, dalam mengembangkan dan meningkatkan kualitas produk berdasarkan preferensi pelanggan

1.5 Batasan Masalah

1. Dataset yang dikumpulkan yaitu ulasan pengguna *e-wallet* dalam Bahasa Indonesia di Google Play Store
2. Dataset di label secara manual menjadi tiga kelas sentimen yaitu positif, netral, dan negatif

1.6 Sistematika Penulisan

BAB I PENDAHULUAN berisi latar belakang, tujuan penelitian, manfaat, penelitian, batasan masalah dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA berisi landasan teori yang digunakan untuk menganalisis masalah yang akan diteliti dan digunakan untuk menyusun tugas akhir ini.

BAB III METODOLOGI PENELITIAN berisi tentang tahapan pada penelitian, teknik pengolahan data, metode penelitian, algoritma penelitian, teknik evaluasi hasil, dan implementasinya.

BAB IV HASIL DAN PEMBAHASAN berisi hasil dan pembahasan dari sistem yang telah dibangun.

BAB V PENUTUP berisi kesimpulan berdasarkan hasil penelitian dan saran untuk pengembangan penelitian selanjutnya

BAB II

TINJAUAN PUSTAKA

2.1 E-Wallet

E-wallet merupakan produk digital yang menawarkan banyak fitur dengan tujuan utama sama seperti dompet saku pada umumnya, bedanya dompet tersebut dalam bentuk aplikasi *mobile*. Awalnya, *e-wallet* dikenal sebagai metode untuk penyimpanan uang elektronik, namun seiring berjalannya waktu serta persaingan antara perusahaan *e-wallet*, tidak hanya digunakan untuk menyimpan uang tetapi untuk media informasi belanja secara daring (*online*) (Basryah et al., 2021).

Bank Indonesia menyatakan bahwa pada tahun 2020 terdapat 47 perusahaan *e-wallet* yang telah memiliki izin resmi (Kristiyanti et al., 2020). Seiring berjalannya waktu semakin banyak produk *e-wallet* yang berkembang dan menawarkan fitur yang bermacam-macam. OVO, Dana, Payfazz, Linkaja dan Isaku adalah beberapa *e-wallet* yang memiliki jumlah unduhan terbanyak di Google Play Store dan populer di masyarakat (Basryah et al., 2021).

2.2 Analisis Sentimen

Analisis Sentimen berkaitan dengan melakukan analisis terhadap emosi dan perspektif penulis dari penggalan teks tertentu. Analisis Sentimen atau *opinion mining* mengacu pada alat pemrosesan bahasa, linguistik, dan analisis teks untuk mencari dan mengekstrak informasi subjektif (Kautish & Kaur, 2017).

Di bidang Natural Language Processing, Analisis Sentimen yang sangat berkembang juga dikenal sebagai ekstraksi perasaan. Ekstraksi perasaan yang

diungkapkan dapat mengendalikan pengambilan keputusan oleh manusia. Analisis Sentimen bertujuan untuk mengklasifikasikan polaritas sebuah teks apakah bersifat positif, negatif maupun netral (Dhuria, 2015). Secara garis besar, tujuan analisis sentimen yaitu untuk mengidentifikasi sikap dari penulis atau polaritas kontekstual terhadap suatu topik berdasarkan teks yang mengekspresikan pendapat (Safina, 2020).

2.3 Bag of Words

Bag of Words (BoW) merupakan representasi sederhana dari yang digunakan pada Natural Language Processing (NLP) dan Information Retrieval (IR). Model BoW mempelajari kosakata dari semua dokumen, dengan menghitung berapa kali sebuah kata muncul. Dalam model ini, sebuah teks (seperti kalimat atau dokumen) direpresentasikan sebagai kantung (*multiset*) dari kata-katanya, dengan mengabaikan tata bahasa dan urutan kata, tetapi tetap mempertahankan keberagamannya (Raj, 2016).

Model BoW umumnya digunakan dalam metode klasifikasi dokumen, dimana frekuensi kemunculan setiap kata digunakan sebagai fitur untuk *training* klasifikasi. Frekuensi setiap kata dalam kosakata akan disimpan. Lalu nilai frekuensi ini tercantum dalam vektor, dimana setiap dimensi mewakili sebuah kata dalam kosakata. Sehingga vektor ini adalah kumpulan kata-kata yang mewakili keseluruhan teks (Schoot Uiterkamp, 2019).

2.4 Algoritma Naïve Bayes

Naïve Bayes merupakan metode yang biasa digunakan untuk kategorisasi dan pelatihan. Naïve Bayes adalah metode klasifikasi berdasarkan probabilitas berdasarkan teorema Bayes untuk memprediksi sekumpulan fitur tertentu sebagai bagian dari label tertentu (Wankhade et al., 2022). Ide inferensi Bayesian ini telah dikenal sejak karya Bayes (1763) dan pertama kali digunakan untuk klasifikasi teks oleh Mosteller and Wallace (1964). Disebut sebagai klasifikasi Bayes yang naif karena klasifikasi ini membuat asumsi penyederhanaan (naif) tentang bagaimana fitur berinteraksi (Jurafsky & Martin, 2019).

Metode Naïve Bayes memiliki keuntungan karena dapat dilatih dan digunakan pada data skala kecil dan dapat diprediksi secara *real-time*. Kelebihan lainnya dapat mengklasifikasikan suatu kelas yang hasilnya dapat digunakan secara paralel dalam meningkatkan skala dataset, terutama pada studi kasus data skala besar (Wongkar & Angdresey, 2019). Algoritma Naïve Bayes memiliki persamaan sebagai berikut

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \quad (1)$$

Dimana

$P(C|X)$ = probabilitas hipotesis C berdasarkan kondisi X (*posterior probability*)

$P(C)$ = probabilitas hipotesis C (*prior probability*)

$P(X|C)$ = probabilitas X berdasarkan kondisi pada hipotesis C

$P(X)$ = probabilitas atribut X

2.5 Multinomial Naïve Bayes

Untuk melakukan klasifikasi dokumen teks, maka algoritma Multinomial Naïve Bayes akan memilih kelas dengan probabilitas tertinggi (Manning et al., 2012), yang dapat dirumuskan dengan persamaan berikut

$$C_{\text{map}} = \arg \max_{c \in C} [\log \hat{P}(c) + \sum_{1 \leq k \leq n_d} \log \hat{P}(t_k | c)] \quad (2)$$

Dimana

C_{map} = kelas dengan probabilitas tertinggi

$\hat{P}(c)$ = *prior probability* kelas c

$\hat{P}(t_k | c)$ = *conditional probability* kelas c

Setiap parameter kondisional $\log \hat{P}(t_k | c)$ adalah bobot yang menunjukkan seberapa baik indikator t_k untuk kelas c . Demikian pula $\log \hat{P}(c)$ adalah bobot yang menunjukkan frekuensi relatif dari kelas c . Kelas yang memiliki frekuensi kemunculan lebih tinggi cenderung menjadi kelas yang benar daripada kelas yang memiliki frekuensi kemunculan rendah. Penjumlahan *log prior* dan bobot kata menjadi tolak ukur untuk melihat seberapa banyak bukti yang ada untuk suatu dokumen di dalam kelas. Persamaan di atas menghitung kelas dengan bukti yang paling banyak.

Kita mencoba menghitung *maximum likelihood estimate* untuk memperkirakan parameter $\hat{P}(c)$ dan $\hat{P}(t_k | c)$, yang merupakan frekuensi relatif dan sesuai dengan nilai yang probabilitasnya paling tinggi dari setiap parameter yang

diberikan dari data latih (Manning et al., 2012). Untuk menghitung *prior* dapat dilakukan dengan menggunakan persamaan

$$\hat{P}(c) = \frac{N_c}{N} \quad (3)$$

Dimana

N_c = jumlah dokumen di dalam kelas c

N = jumlah keseluruhan dokumen

Lalu menghitung *conditional probability* sebagai frekuensi relative dari *term* t di dalam dokumen kelas c dengan persamaan

$$\hat{P}(t|c) = \frac{T_{ct}+1}{\sum_{t \in V} (T_{ct'}+1)} = \frac{T_{ct}+1}{(\sum_{t \in V} T_{ct'})+B'} \quad (4)$$

Dimana

T_{ct} = jumlah kemunculan t dalam dokumen pelatihan dari kelas c

$B = |V|$ = jumlah kata unik pada keseluruhan dokumen

Karena kita mengalikan semua *conditional probability* ke dalam persamaan (2), maka untuk menghindari hasil probabilitas sama dengan 0 pada kata yang tidak ada dalam suatu kelas maka, digunakan *add-one* dengan menambahkan 1 atau *Laplace smoothing* ke dalam persamaan (4).

2.6 Optimasi Naïve Bayes Untuk Analisis Sentimen

Salah satu tambahan yang penting dalam klasifikasi teks untuk sentimen adalah menangani negasi. Hal sederhana yang dapat dilakukan dalam analisis sentimen untuk menangani negasi adalah pada normalisasi teks dengan

menambahkan prefix “NOT_” ke setiap kata setelah tanda negasi (tidak) hingga tanda baca berikutnya (Jurafsky & Martin, 2019). Sebagai contoh,

saya tidak suka film ini, tapi saya

menjadi

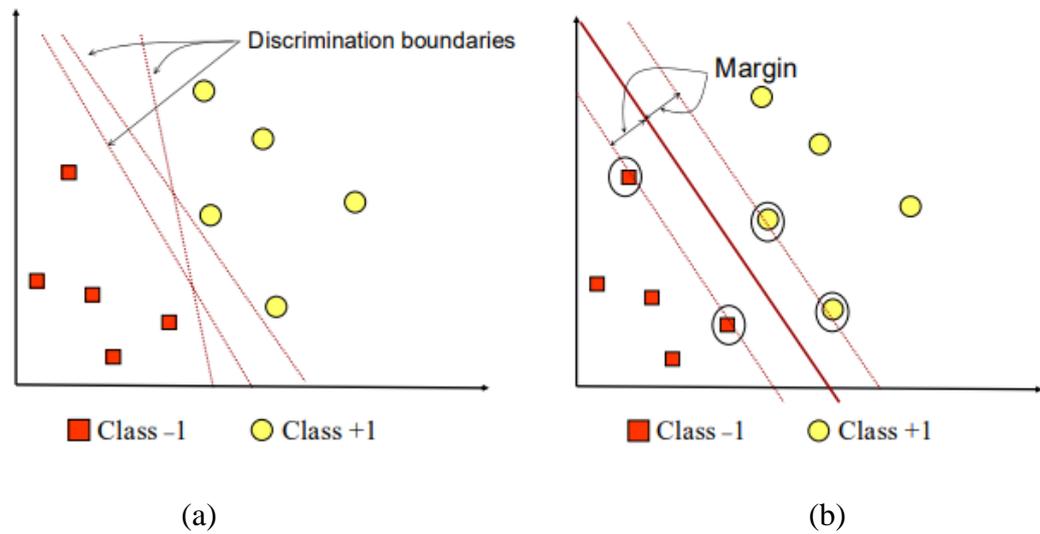
saya tidak NOT_suka NOT_film NOT_ini, tapi saya

Kata-kata baru yang terbentuk seperti NOT suka atau NOT rekomendasi akan lebih sering muncul pada kelas dokumen negatif dan bertindak sebagai isyarat dokumen negatif. Sementara untuk kata seperti NOT bosan atau NOT menolak masuk ke dalam kelas dokumen positif (Jurafsky & Martin, 2019).

2.7 Support Vector Machine

Support vector Machine (SVM) adalah metode klasifikasi linear. Peran SVM dalam klasifikasi yaitu dengan cara menentukan *separator* atau garis pembatas yang disebut sebagai *hyperplane* di ruang pencarian yang membatasi antara kelas yang berbeda. Salah satu keunggulan dari SVM yaitu sangat baik dalam klasifikasi data dengan dimensional tinggi karena metode ini menentukan arah diskriminan yang optimal dalam ruang fitur dengan cara menemukan kombinasi fitur yang paling tepat (Yonatha Wijaya & Karyawati, 2020).

Garis solid pada Gambar 2.7.1 (b) menunjukkan *hyperplane* yang terbaik, yaitu yang terletak tepat di tengah anantara kedua kelas, sedangkan titik merah dan kuning yang berada dalam lingkaran hitam adalah *support vector* (Nugroho et al., 2005).



Gambar 2.7.1 Hyperplane antara kelas positif dan negatif (Nugroho et al., 2005)

Secara matematika, formulasi problem optimisasi SVM untuk kasus klasifikasi linier Pegasos adalah (Shalev-Shwartz et al., 2000),

$$\min_w \frac{\lambda}{2} \|w\|^2 + \frac{1}{m} \sum_{(x,y) \in S} \ell(w; +(x, y)) \quad (5)$$

Dimana

$$\ell(w; +(x, y)) = \max \{0, 1 - y(w, x)\} \quad (6)$$

Diberikan set data latih $\{(x_i, y_i)\}_{i=1}^m$, dimana x_i adalah data input, y_i adalah output yang nilainya 1 atau -1. Dengan formulasi di atas digunakan untuk meminimalkan fungsi tujuan (*objective function*).

Inisialnya, bobot vektor diatur menjadi $w_1 = 0$. Untuk setiap iterasi t pada algoritma, yang pertama akan dilakukan adalah memilih data latih (x_{i_t}, y_{i_t}) pada indeks $i_t \in \{1, \dots, m\}$ secara acak. Lalu mengganti fungsi dalam persamaan (5) dengan perkiraan berdasarkan data latih (x_{i_t}, y_{i_t}) yang menghasilkan,

$$f(w; i_t) = \frac{\lambda}{2} \|w\|^2 + \ell(w; x_{i_t}, y_{i_t}) \quad (7)$$

Dengan mempertimbangkan *sub-gradien* dari fungsi tujuan di atas, maka persamaan menjadi,

$$\nabla_t = \lambda w_t - \mathbb{1}[y_{i_t}(w_t, x_{i_t}) < 1] y_{i_t} x_{i_t} \quad (8)$$

Dimana $\mathbb{1}[y_{i_t}(w_t, x_{i_t})]$ merupakan fungsi indikator yang mengambil nilai satu jika argumennya benar. Kemudian diperbarui persamaan $w_{t+1} \leftarrow w_t - \eta_t \nabla_t$ menggunakan variabel *step size* $\eta_t = \frac{1}{\lambda t}$. Sehingga pembaharuan ini menjadi persamaan berikut (Shalev-Shwartz et al., 2000)

$$w_{t+1} \leftarrow \left(1 - \frac{1}{t}\right) w_t + \eta_t \mathbb{1}[y_{i_t}(w_t, x_{i_t}) < 1] y_{i_t} x_{i_t} \quad (9)$$

Cara kerja SVM adalah teknik *binary classification*, dimana nilai -1 menunjukkan satu kelas dan +1 menunjukkan kelas lainnya. Maka persamaan klasifikasi linearnya (Shiri, 2004), yaitu

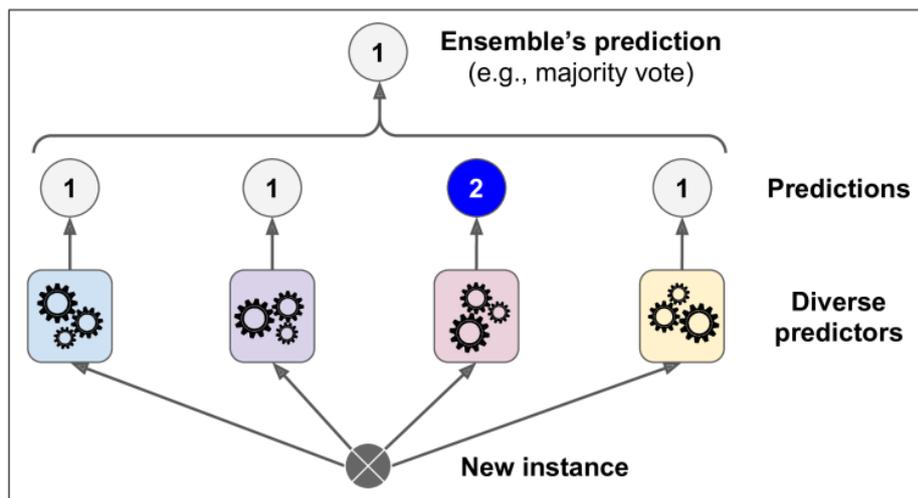
$$f(x) = \text{sign}(w_t x) \quad (10)$$

Oleh karena itu untuk membangun *multi-class classification* maka teknik *One-Versus-All* (OVA) digunakan. Dimana dengan metode ini masalah klasifikasi kelas-kelas N dipecah menjadi kelas N masalah biner. Selama proses *training* sampel kelas ke-n diberi label positif (+) dan sisa contoh N-1 diberi label negatif (-) (Żurek & Pietronó, 2020). Sehingga dapat menghasilkan persamaan berikut

$$f_{\text{ova}}(x) = \text{argmax}_{i=1 \dots N} f_i(x) \quad (11)$$

2.8 Ensemble Learning

Ensemble Learning Model adalah prosedur yang digunakan untuk melatih banyak model machine learning dan menggabungkan outputnya untuk mendapatkan model global komposit yang lebih baik dengan keputusan yang lebih akurat dan andal daripada yang dicapai melalui satu model (Ali et al., 2015).



Gambar 2.8.1 Ensembl Hard Voting (Géron, 2019)

Ada dua jenis Voting Ensemble yang cukup populer, yaitu Hard Voting dan Soft Voting (Peppes et al., 2021). Hard voting bekerja sangat sederhana dengan menggabungkan prediksi setiap model klasifikasi dan mengambil kelas prediksi dengan suara terbanyak (Géron, 2019). Sedangkan Soft Voting adalah metode yang lebih kompleks yang memperhitungkan probabilitas setiap prediksi oleh masing-masing model klasifikasi dasar (Peppes et al., 2021).

2.9 Confusion Matrix

Di data mining khususnya dalam klasifikasi teks, kita membutuhkan metrik untuk mengetahui seberapa baik metode dan sistem yang dibuat. Confusion Matrix

dapat digunakan untuk merangkum kinerja model klasifikasi terhadap data uji. Confusion Matrix adalah tabel untuk memvisualisasikan bagaimana kinerja suatu algoritma terhadap kelas aktual, menggunakan dua dimensi yaitu kelas prediksi dan kelas aktual, dan setiap sel memberi kemungkinan satu set hasil. (Jurafsky & Martin, 2019).

		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

(a)

		Predicted Class			
		C_1	C_2	...	C_N
Actual Class	C_1	$C_{1,1}$	FP	...	$C_{1,N}$
	C_2	FN	TP	...	FN

	C_N	$C_{N,1}$	FP	...	$C_{N,N}$

(b)

Gambar 2.9.1 Confusion matrix: (a) *Binary classification* (b) *Multiclass classification* (Markoulidakis et al., 2021)

Confusion Matrix untuk masalah *binary classification* dan *multiclass classification* ditampilkan pada Gambar 2.9.1. Setiap kolom matriks mewakili *instance* dari kelas yang diprediksi, sedangkan setiap baris mewakili *instance* dari kelas yang sebenarnya. Sebuah elemen dari Confusion Matrix pada baris i dan kolom j memberikan jumlah kejadian dimana kelas yang diprediksi adalah j dan kelas yang sebenarnya adalah i . Dalam konteks ini, Confusion Matrix menunjukkan cara model klasifikasi yang "bingung" ketika membuat prediksi. Ini dapat memberikan wawasan tidak hanya tentang kesalahan yang dibuat oleh pengklasifikasi, tetapi yang lebih penting adalah jenis kesalahan yang dihasilkan (Markoulidakis et al., 2021).

Untuk mengukur akurasi dari sebuah model juga dapat dilakukan dengan menggunakan Confusion Matrix. Begitu juga dengan mengukur Recall atau *True Positive Rate*, dan Precision atau *Positive Predictive Value* (Markoulidakis et al., 2021). Berikut persamaannya

Akurasi

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (12)$$

Recall

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (13)$$

Precision

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (14)$$

Dimana

TP (*True Positive*) = jumlah kelas positif yang diklasifikasi sebagai kelas positif

TN (*True Negative*) = jumlah kelas negatif yang di klasifikasi sebagai kelas negatif

FP (*False Positive*) = jumlah kelas negatif yang di klasifikasi sebagai kelas positif

FN (*False Negative*) = jumlah kelas positif yang diklasifikasi sebagai kelas negatif

Akurasi mungkin adalah salah satu metrik paling dasar, tetapi akurasi tidak digunakan secara umum untuk klasifikasi teks. Ini karena akurasi tidak bekerja dengan baik apabila jumlah data setiap kelas tidak seimbang. Sehingga biasanya Precision dan Recall menjadi dua metrik yang digunakan untuk mengevaluasi

model. Adapun F-measure menjadi salah satu metrik yang digunakan untuk menggabungkan aspek dari keduanya yang menggunakan parameter β untuk membedakan bobot recall dan precision, yang didefinisikan sebagai (Jurafsky & Martin, 2019):

$$F_{\beta} = \frac{(\beta^2+1)PR}{\beta^2P+R} \quad (15)$$

Jika nilai $\beta > 1$ maka mendukung recall, sedangkan jika $\beta < 1$ maka mendukung precision. Ketika $\beta = 1$, dimana precision dan recall seimbang, maka berikut adalah persamaan $F1$ yang biasanya digunakan:

$$F_1 = \frac{2PR}{P+R} \quad (16)$$