SKRIPSI

PERBANDINGAN METODE KLASIFIKASI RANDOM FOREST, XGBOOST DAN SVM PADA ANALISIS SENTIMEN APLIKASI KREDIT DAN PINJAMAN ONLINE

Disusun dan diajukan oleh:

PAHRUL
D121 19 1020

PROGRAM STUDI SARJANA TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS HASANUDDIN
GOWA
2023

LEMBAR PENGESAHAN SKRIPSI

PERBANDINGAN METODE KLASIFIKASI RANDOM FOREST, XGBOOST DAN SVM PADA ANALISIS SENTIMEN APLIKASI KREDIT DAN PINJAMAN ONLINE

Disusun dan diajukan oleh

PAHRUL
D121191020

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Studi Program Sarjana Program Studi Teknik Informatika
Fakultas Teknik Universitas Hasanuddin
Pada tanggal 7 Juni 2023
dan dinyatakan telah memenuhi syarat kelulusan

Menyetujui,

Pembimbing Utama,	Pembimbing Pendamping,
Dr. Amil Ahmad Ilham, S.T., M.IT.	Elly Warni, S.T., M.T.
NIP 197310101998021001	NIP 198202162008122001

Ketua Program Studi,

Prof. Dr. Ir. Indrabayu, S.T., M.T.,M.Bus.Sys., IPM., ASEAN.Eng.
NIP 197507162002121004

PERNYATAAN KEASLIAN

Yang bertanda tangan dibawah ini:

Nama	:	Pahrul
NIM	:	D121191020
Program Studi	:	Teknik Informatika
Jenjang	:	S1

Menyatakan dengan ini bahwa karya tulisan saya berjudul

Perbandingan Metode Klasifikasi Random Forest, XGBoost dan SVM pada Analisis Sentimen Aplikasi Kredit dan Pinjaman Online

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

Semua informasi yang ditulis dalam skripsi yang berasal dari penulis lain telah diberi penghargaan, yakni dengan mengutip sumber dan tahun penerbitannya. Oleh karena itu semua tulisan dalam skripsi ini sepenuhnya menjadi tanggung jawab penulis. Apabila ada pihak manapun yang merasa ada kesamaan judul dan atau hasil temuan dalam skripsi ini, maka penulis siap untuk diklarifikasi dan mempertanggungjawabkan segala resiko.

Segala data dan informasi yang diperoleh selama proses pembuatan skripsi, yang akan dipublikasi oleh Penulis di masa depan harus mendapat persetujuan dari Dosen Pembimbing.

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

Gowa, 11 Juni 2023

Yang Menyatakan

Pahrul

ABSTRAK

PAHRUL. *Perbandingan Metode Klasifikasi Random Forest, XGBoost dan SVM pada Analisis Sentimen Aplikasi Kredit dan Pinjaman Online* (dibimbing oleh Amil Ahmad Ilham dan Elly Warni)

Aplikasi pinjaman *online* telah memungkinkan masyarakat untuk memperoleh pinjaman dengan mudah. Namun, seperti halnya dengan layanan keuangan lainnya, penggunaan pinjaman *online* memiliki risiko. Beberapa risiko yang mungkin terjadi antara lain penggunaan yang tidak bertanggung jawab, penipuan, biaya yang terlalu tinggi, dan terjebak dalam siklus hutang. Oleh karena itu, pengguna harus memahami risiko-risiko ini dan melakukan penilaian risiko sebelum memutuskan untuk mengambil pinjaman *online* dari perusahaan yang tepat. Sistem yang dapat dikembangkan untuk memberikan rekomendasi aplikasi pinjaman *online* adalah melalui analisis sentimen dengan memanfaatkan ulasan atau penilaian masyarakat terhadap aplikasi tersebut. Penelitian ini menggunakan metode klasifikasi Random Forest, XGBoost, dan Support Vector Machine untuk mengembangkan sistem dan menggabungkan hasil dari ketiga model tersebut menggunakan metode *ensemble learning* Soft Voting Classifier untuk memperoleh akurasi yang lebih tinggi. Dari hasil pengujian, model Soft Voting memiliki performa yang unggul dibandingkan dengan model Random Forest, XGBoost, dan Support Vector Machine dengan akurasi sebesar 87.3%. Model Random Forest menghasilkan akurasi sebesar 80.9%, disusul oleh model XGBoost dengan akurasi sebesar 80.7%, dan Support Vector Machine menjadi model dengan performa terendah dengan akurasi sebesar 79.9%. Oleh karena itu, diputuskan menggunakan model Soft Voting Classifier yang memiliki tingkat akurasi tertinggi untuk mengklasifikasikan sentimen ulasan pengguna untuk merekomendasikan aplikasi pinjaman *online*.

Kata Kunci: Random Forest, XGBoost, Support Vector Machine, Soft Voting, Analisis Sentimen

ABSTRACT

PAHRUL. *Comparison of Random Forest, XGBoost and SVM Classification Methods on Sentiment Analysis of Online Credit and Loan Applications* (supervised by Amil Ahmad Ilham and Elly Warni).

Online loan applications have enabled people to obtain loans easily. However, like other financial services, online loans carry risks. Some possible risks include irresponsible usage, fraud, high fees, and being trapped in a debt cycle. Therefore, users must understand these risks and assess these risks to take out online loans from the right company. A system that can be developed to provide recommendations for online loan applications is through sentiment analysis by utilizing reviews or ratings from the public on the application. This study uses the classification methods of Random Forest, XGBoost, and Support Vector Machine to develop a system and combines the results from all three models using the Soft Voting Classifier ensemble learning method to achieve higher accuracy. From the testing results, the Soft Voting model performed better than the Random Forest, XGBoost, and Support Vector Machine models with an accuracy of 87.3%. The Random Forest model produced an accuracy of 80.9%, followed by the XGBoost model with an accuracy of 80.7%, and Support Vector Machine had the lowest performance with an accuracy of 79.9%. Therefore, using the Soft Voting Classifier model with the highest accuracy rate was decided to classify user review sentiments for recommending online loan applications.

Keywords: Random Forest, XGBoost, Support Vector Machine, Soft Voting, Sentiment Analysis.

DAFTAR ISI

ABSTRAK ii
ABSTRACT iii
DAFTAR ISI iv
DAFTAR GAMBAR vi
DAFTAR TABEL vii
DAFTAR SINGKATAN DAN ARTI SIMBOL viii
DAFTAR LAMPIRAN ix
KATA PENGANTAR x
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 3
1.3 Tujuan Penelitian 4
1.4 Manfaat Penelitian 4
1.5 Batasan Masalah 4
1.6 Sistematika Penulisan 4
BAB II TINJAUAN PUSTAKA 6
2.1 Aplikasi Kredit dan Pinjaman *Online* 6
2.2 Kredivo 7
2.3 Indodana 7
2.4 AdaKami 7
2.5 Akulaku 8
2.6 *Natural Language Processing* 8
2.7 Analisis Sentimen 9
2.8 Word2Vec 10
2.9 Random Forest 12
2.10 eXtreme Gradient Boosting (XGBoost) 14
2.11 Support Vector Machine (SVM) 16
2.12 Particle Swarm Optimization 19
2.13 *Ensemble Learning* 19
2.14 Confusion Matrix 21
BAB III METODE PENELITIAN 26
3.1 Tempat dan Waktu 26
3.2 Alat dan Bahan 26
3.3 Tahapan Penelitian 26
3.4 Rancangan Sistem 29
3.5 Pengumpulan Data 30
3.6 Pelabelan Data 30
3.8 Representasi Kata 35
3.9 Model Klasifikasi Random Forest 36
3.10 Model Klasifikasi XGBoost 37
3.11 Model Klasifikasi Support Vector Machine (SVM) 39
3.12 Optimasi Bobot Menggunakan PSO 41
3.13 Implementasi Soft Voting Classifier 42
BAB IV HASIL DAN PEMBAHASAN 45

4.1 Pengumpulan Data 45
4.2 Pelabelan Data 45
4.3 *Preprocessing* Data 46
4.4 Analisis Sentimen dengan Random Forest 49
4.5 Analisis Sentimen dengan XGBoost 51
4.6 Analisis Sentimen dengan Support Vector Machine 52
4.7 Analisis Sentimen dengan Soft Voting Classifier 54
4.8 Perbandingan Kinerja Model 56
4.9 Prediksi Sentimen dan Rekomendasi Aplikasi 57
BAB V KESIMPULAN DAN SARAN 65
5.1 Kesimpulan 65
5.2 Saran 65
DAFTAR PUSTAKA 67
LAMPIRAN 71

DAFTAR GAMBAR

Gambar 2.8.1 Model CBOW (Rong, 2016) 11
Gambar 2.8.2 Model Skip-gram (Rong, 2016) 12
Gambar 2.9.1 Cara Kerja Algoritma Random Forest (Al Amrani et al., 2018) 13
Gambar 2.10.1 Diagram Skema XGBoost (Guo et al., 2020) 15
Gambar 2.11.1 *Hyperplane* (Nugroho, 2008) 17
Gambar 2.12.1 Soft Voting Classifier (Manconi et al., 2022) 20
Gambar 3.3.1 Tahapan Penelitian 27
Gambar 3.4.1 Rancangan Sistem 29
Gambar 3.9.1 Diagram Proses Random Forest 36
Gambar 3.10.1 Diagram Proses XGBoost 37
Gambar 3.11.1 Diagram Proses SVM 39
Gambar 3.12.1 Flowchart Optimasi Bobot 41
Gambar 3.13.1 Soft Voting Classifier 43
Gambar 4.1.1 Contoh Hasil *Scrape* Dataset 45
Gambar 4.2.1 Grafik Pembagian Kelas Dataset 46
Gambar 4.4.1 Confusion Matrix Random Forest 50
Gambar 4.5.1 Confusion Matrix XGBoost 51
Gambar 4.6.1 Confusion Matrix SVM 53
Gambar 4.7.1 Confusion Matrix Soft Voting 54
Gambar 4.8.1 Grafik Perbandingan Kinerja Model 57
Gambar 4.9.1 Jumlah Sentimen Kredivo 58
Gambar 4.9.2 Grafik Jumlah Sentimen Kredivo Berdasarkan Tanggal 58
Gambar 4.9.3 Jumlah Sentimen Indodana 59
Gambar 4.9.4 Grafik Jumlah Sentimen Indodana Berdasarkan Tanggal 60
Gambar 4.9.5 Jumlah Sentimen AdaKami 61
Gambar 4.9.6 Grafik Jumlah Sentimen AdaKami Berdasarkan Tanggal 61
Gambar 4.9.7 Jumlah Sentimen Akulaku 62
Gambar 4.9.8 Grafik Jumlah Sentimen Akulaku Berdasarkan Tanggal 63

DAFTAR TABEL

Tabel 2.14.1 Confusion Matrix *Multiclass* 21
Tabel 2.14.2 Nilai *False Positive* 22
Tabel 2.14.3 Nilai *True Negative* Kelas Netral 23
Tabel 2.14.4 Nilai *True Negative* Kelas Negatif 23
Tabel 2.14.5 Nilai *True Negative* Kelas Positif 24
Tabel 2.14.6 Nilai *False Negative* 24
Tabel 3.7.1 Contoh Kamus Normalisasi Kata 34
Tabel 3.8.1 Representasi Numerik Kata Menggunakan Word2Vec 35
Tabel 4.2.1 Rincian Jumlah Ulasan pada Setiap Kelas 46
Tabel 4.3.1 Ulasan Sebelum dan Sesudah Proses *Remove Punctuation* 46
Tabel 4.3.2 Ulasan Sebelum dan Sesudah Proses *Case Folding* 47
Tabel 4.3.3 Ulasan Sebelum dan Sesudah Proses *Tokenizing* 47
Tabel 4.3.4 Ulasan Sebelum dan Sesudah Proses *Stopword Removal* 48
Tabel 4.3.5 Ulasan Sebelum dan Sesudah Proses Normalisasi 48
Tabel 4.3.6 Ulasan Sebelum dan Sesudah Proses *Stemming* 49
Tabel 4.4.1 Evaluasi Model Random Forest 50
Tabel 4.5.1 Evaluasi Model XGBoost 52
Tabel 4.6.1 Evaluasi Model SVM 53
Tabel 4.7.1 Evaluasi Model Soft Voting 55
Tabel 4.8.1 Perbandingan Kinerja Model 56

DAFTAR SINGKATAN DAN ARTI SIMBOL

Lambang/Singkatan	Arti dan Keterangan
XGBoost	*eXtreme Gradient Boosting*
SVM	*Support Vector Machine*
NLP	*Natural Language Processing*
CBOW	*Continuous Bag-of-Words*
OVA	*One-vs-All*
AUC	*Area Under the Curve*
Fintech	*Financial Technology*

DAFTAR LAMPIRAN

Lampiran 1. Hasil Prediksi Model 71
Lampiran 2. *Preprocessing* Data 72
Lampiran 3. *Class* Random Forest 74
Lampiran 4. *Class* XGBoost 75
Lampiran 5. *Class* Support Vector Machine 77
Lampiran 6. Optimasi Bobot 80
Lampiran 7. Contoh Sederhana Cara Kerja SVM Menghasilkan Prediksi 81
Lampiran 8. Contoh Sederhana Cara Kerja Word2Vec 84
Lampiran 9. Contoh Sederhana Cara Kerja Random Forest 88
Lampiran 10. Contoh Sederhana Cara Kerja XGBoost 92
Lampiran 11. Skenario Performa Baru dari Penambahan Jumlah Data 95

KATA PENGANTAR

Puji dan syukur penulis panjatkan kehadirat Allah SWT karena atas berkat dan rahmat-Nya sehingga dapat menyelesaikan tugas akhir dengan judul "**Perbandingan Metode Klasifikasi Random Forest, XGBoost dan SVM pada Analisis Sentimen Aplikasi Kredit dan Pinjaman *Online***" sebagai salah satu persyaratan akademik untuk menyelesaikan program Strata-1 pada 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 atas berkat dan rahmat-Nya sehingga penulis dapat menyelesaikan tugas akhir ini.
2. Kedua orang tua penulis, Bapak Asri yang senantiasa mendoakan dan memberikan dukungan dalam menyelesaikan perkuliahan dan Ibu Ernawati, yang selalu menjadi tempat curhat saya dan selalu siap untuk memfasilitasi penulis dalam menjalankan dunia perkuliahan. Terima kasih atas kesabaran dan dukungannya selama ini.
3. Bapak Dr. Amil Ahmad Ilham, S.T., M.IT. selaku pembimbing I dan Ibu Elly Warni, ST., 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. Alifa Nur Fadila, Afifah Mardhiyah Ramlan, dan Lutfiah Salim, yang telah menjadi teman yang baik dan menyenangkan sejak masa-masa awal perkuliahan. Kebersamaan kita dalam mengerjakan tugas-tugas kuliah, diskusi, dan juga sekedar menghabiskan waktu luang di kampus akan selalu menjadi kenangan yang berharga bagi penulis.

6. Ariyanti Herlota dan Alya Ramadani, dua sosok yang telah senantiasa memberikan semangat kepada penulis. Kata-kata terima kasih tak akan pernah cukup untuk mengungkapkan rasa terima kasih penulis kepada kalian berdua.
7. Deby Rizky Ramadana, Farhan Adyatama, Andi Rusmiati, Dea Wahsa, nur faiz ramadhan, Wira Drana Wasistha dan Leonic yang senantiasa menghibur penulis dalam menyelesaikan tugas akhir dan senantiasa meluangkan tenaga dan waktunya untuk mendengarkan curhatan dari penulis.
8. Teman-teman Teknik Informatika Angkatan 2019 (S19NIFIER) khususnya ketua angkatan (Giga), teman-teman informatika kelas A, dan juga terkhusus kepada teman-teman lab (Dita, Sayid, Rayyan dan Juan) yang telah memberi bantuan, dukungan dan semangat selama masa perkuliahan dan penyusunan tugas akhir ini.
9. Teman-teman KKN Posko Desa Tunikamaseang Gel.108 (Isra, Asep, Arsyi, Alief, Uul, Shani, Cindy, Ima, Ila, Nia dan Liani) yang telah memberi pengalaman berkesan kepada penulis selama KKN.
10. Serta berbagai pihak atas segala dukungan dan bantuannya yang tidak dapat penulis tuliskan satu persatu.

Penulis berharap semoga Tuhan 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.

Gowa, Juni 2023

Penulis,

Pahrul

BAB I
PENDAHULUAN

1.1 Latar Belakang

Financial Technology atau *fintech* adalah istilah yang merujuk pada inovasi dalam layanan keuangan yang menggunakan teknologi untuk meningkatkan kinerja, efisiensi, dan kemudahan akses. *Fintech* hadir seiring dengan perubahan gaya hidup masyarakat yang kini banyak menggunakan teknologi informasi untuk memenuhi tuntutan hidup yang semakin cepat. Pengaruh *fintech* terhadap masyarakat luas sangat signifikan, karena memberikan akses yang lebih mudah terhadap produk keuangan, memungkinkan transaksi dilakukan secara praktis dan efektif melalui smartphone, *e-Money*, bahkan investasi. Permasalahan dalam transaksi seperti sulitnya mencari barang di tempat perbelanjaan, kesulitan mentransfer dana ke bank atau ATM, atau enggan berkunjung ke tempat tertentu karena pelayanan yang tidak menyenangkan, dapat diminimalkan dengan *fintech* sehingga semuanya dapat dilakukan dengan mudah (Kamil et al., 2022).

Salah satu produk *financial technology* atau *fintech* di Indonesia adalah pinjaman *online* dalam bentuk aplikasi. Pinjaman *online* dalam bentuk aplikasi mampu membuat masyarakat memperoleh pinjaman dengan mudah, hal tersebut membuat banyak masyarakat tertarik untuk melakukan pinjaman *online* di aplikasi sehingga seiring dengan meningkatnya minat masyarakat terhadap pinjaman *online* membuat penyedia layanan *financial technology* semakin banyak dan membuat persaingan antara para perusahaan penyedia *financial technology* dalam menarik minat para masyarakat.

Pinjaman *online* adalah layanan keuangan yang memungkinkan pengguna untuk meminjam uang melalui platform digital. Biasanya, layanan pinjaman *online* ini menawarkan persyaratan yang lebih fleksibel daripada layanan pinjaman tradisional, dan pengguna dapat mengajukan pinjaman dengan cepat dan mudah tanpa perlu mengunjungi kantor bank atau lembaga keuangan lainnya. Salah satu alasan utama mengapa orang memilih pinjaman online adalah karena kebutuhan mendesak yang harus dipenuhi. Terkadang, ada keadaan darurat di mana seseorang membutuhkan dana segera, seperti membayar biaya kesehatan

yang tidak terduga, kebutuhan darurat di tengah bulan, atau membayar tagihan listrik atau air yang harus dibayar dalam waktu singkat. Namun, seperti halnya dengan layanan keuangan lainnya, ada risiko terkait dengan penggunaan pinjaman *online*. Beberapa risiko yang dapat terjadi adalah penggunaan yang tidak bertanggung jawab, penipuan, biaya yang terlalu tinggi, dan terjebak dalam siklus hutang. Oleh karena itu, pengguna perlu memahami risiko-risiko ini dan melakukan penilaian risiko sebelum memutuskan untuk mengambil pinjaman *online*.

Saat ini, terdapat beberapa aplikasi *online* yang menyediakan kredit dan pinjaman yang sudah dikenal oleh masyarakat seperti Akulaku, Kredivo, Indodana, dan AdaKami. Namun, banyaknya ulasan atau opini yang ada mengenai aplikasi pinjaman *online* ini membuat calon konsumen merasa ragu dan tidak yakin dalam memilih aplikasi mana yang mudah, aman, dan nyaman digunakan, serta yang sesuai dengan kebutuhan dan dapat memberikan kepuasan. Oleh karena itu, diperlukan sebuah sistem yang mampu memberikan saran kepada masyarakat tentang aplikasi pinjaman *online* yang tepat.

Salah satu sistem yang dapat dikembangkan adalah melalui analisis sentimen dengan memanfaatkan ulasan atau penilaian masyarakat terhadap aplikasi pinjaman *online* tersebut. Analisis sentimen adalah teknik pemrosesan bahasa alami (*Natural Language Processing*) yang digunakan untuk mengidentifikasi dan mengekstraksi informasi dari teks berbahasa manusia. Analisis sentimen adalah salah satu penelitian yang sering dimanfaatkan dalam membantu menganalisis emosi, keinginan, dan persepsi seseorang terhadap suatu entitas. Tujuan utama dari analisis sentimen adalah untuk menganalisis pemikiran, mendefinisikan perasaan, dan menentukan polaritas (Sukhavasi, 2021). Teknik ini umumnya digunakan untuk menganalisis perasaan, sikap, atau opini seseorang terhadap suatu topik atau entitas tertentu, seperti produk, merek, atau layanan.

Analisis sentimen dapat dilakukan dengan menggunakan algoritma *machine learning* untuk mengklasifikasikan teks ke dalam kategori sentimen positif, negatif, atau netral berdasarkan kata-kata dan kalimat yang digunakan dalam teks tersebut. Sumber data untuk mendapatkan dataset dalam penelitian analisis

sentimen adalah bisa melalui twitter atau melalui ulasan pelanggan yang ada pada Google Playstore.

Penelitian yang dilakukan oleh Özçift (Özçift, 2020) menggunakan metode Soft Voting untuk melakukan analisis sentimen didapatkan hasil akurasi sebesar 91,2%. Sistem yang dibangun oleh Yassine dengan menggunakan algoritma Random Forest dan Support Vector Machine untuk analisis sentimen didapatkan hasil bahwa Random Forest mampu melakukan prediksi sentimen dengan akurasi sebesar 81% dan jika menggunakan Support Vector Machine didapatkan hasil akurasi sebesar 82.4% yang berarti sebanyak 82.4% di antaranya berhasil diprediksi dengan benar oleh model SVM (Al Amrani et al., 2018). Zongmin Li pada tahun 2022 melakukan penelitian analisis sentimen dengan menggunakan metode klasifikasi XGBoost menghasilkan nilai AUC dengan rata-rata 0.7536 dan nilai F1 dengan rata-rata 0.7490 (Li et al., 2020).

Penelitian ini akan mengembangkan sistem serupa dengan menggunakan metode klasifikasi Random Forest, XGBoost dan Support Vector Machine kemudian menggabungkan hasil dari ketiga model tersebut menggunakan metode *ensemble learning* Soft Voting Classifier. Hasil klasifikasi dari keempat model yang menghasilkan output baik dapat digunakan untuk memberikan daftar rekomendasi aplikasi pinjaman *online*.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah dijelaskan di atas maka rumusan masalah dalam penelitian ini adalah:

1. Bagaimana cara membangun sistem untuk analisis sentimen dengan menggunakan metode klasifikasi Random Forest, XGBoost, Support Vector Machine, dan metode *ensemble* Soft Voting?
2. Bagaimana cara mendapatkan rekomendasi aplikasi kredit dan pinjaman *online* berdasarkan sentimennya?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Membangun sistem dan mengidentifikasi metode yang memiliki performa terbaik.
2. Memberikan rekomendasi aplikasi kredit dan pinjaman *online*.

1.4 Manfaat Penelitian

Penelitian ini diharapkan dapat membantu:

1. Masyarakat, penelitian ini diharapkan dapat dijadikan sumber informasi oleh masyarakat untuk dapat lebih selektif dalam memilih aplikasi kredit dan pinjaman online yang lebih terpercaya,
2. Perusahaan, penelitian ini diharapkan dapat menjadi referensi bagi perusahaan dalam mengembangkan fitur aplikasi dan meningkatkan kualitas pelayanan sesuai dengan ulasan atau review pengguna.
3. Peneliti, penelitian ini diharapkan dapat menjadi referensi bagi akademisi dan peneliti di masa depan yang tertarik untuk melakukan penelitian seputar analisis sentimen atau data mining pada institusi Pendidikan.

1.5 Batasan Masalah

Batasan Masalah dari penelitian ini adalah:

1. Dataset diperoleh dari *scraping* data pada google play store
2. Aplikasi yang akan direview yaitu Akulaku, Kredivo, Indodana, dan AdaKami.
3. Metode klasifikasi yang digunakan adalah Random Forest, XGBoost, dan Support Vector Machine.
4. Metode *ensemble learning* yang digunakan adalah Soft Voting Classifier.
5. Dataset yang digunakan menggunakan bahasa Indonesia.

1.6 Sistematika Penulisan

Untuk menjelaskan secara keseluruhan isi dari laporan tugas akhir ini, berikut dijelaskan bagaimana sistematika penulisan dilakukan.

BAB I PENDAHULUAN

Bagian ini memuat penjelasan mengenai latar belakang, masalah yang dirumuskan, tujuan penelitian, manfaat penelitian, batasan masalah, dan juga sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Bagian ini memuat beberapa referensi literatur yang berkaitan dengan topik penelitian dan teori-teori yang mendukung penelitian yang dilakukan.

BAB III METODOLOGI PENELITIAN

Bagian ini memberikan gambaran tentang desain dan alur sistem serta berisi tentang metode-metode yang akan digunakan dalam penelitian.

BAB IV HASIL DAN PEMBAHASAN

Bagian ini menggambarkan hasil penelitian serta membahas dan mengevaluasi penelitian secara keseluruhan.

BAB V PENUTUP

Bagian ini menyajikan simpulan dari hasil penelitian serta memberikan saran untuk pengembangan sistem di masa depan.

BAB II
TINJAUAN PUSTAKA

2.1 Aplikasi Kredit dan Pinjaman *Online*

Pinjaman *online* adalah sebuah jenis pinjaman yang dilakukan secara daring menggunakan ponsel atau perangkat sejenisnya tanpa harus melakukan tatap muka secara langsung. Pengajuan kredit dan pinjaman yang selama ini dikenal rumit dan memakan waktu yang lumayan lama, sekarang bisa dilakukan secara mudah, cepat dan bisa diajukan tanpa harus tatap muka melalui aplikasi. Pinjaman *online* merupakan salah satu bentuk penggunaan teknologi dalam sistem keuangan yang dapat menghasilkan produk, layanan, teknologi, dan/atau model bisnis baru. Selain itu, pinjaman *online* juga berdampak pada stabilitas moneter, stabilitas sistem keuangan, efisiensi, kelancaran, keamanan, dan keandalan sistem pembayaran (Arista & Rusmini, 2022).

Pada tahun 2017, sebuah surat kabar *online* melaporkan bahwa terdapat sekitar 28 perusahaan yang menawarkan layanan pinjaman *online* dan terdaftar di OJK. Hal ini berarti bahwa terdapat 28 aplikasi pinjaman *online* yang tersedia saat itu. Namun, pada bulan desember tahun 2018, jumlah tersebut meningkat secara signifikan hingga delapan kali lipat menjadi 88 perusahaan *online* yang terdaftar di OJK (Widjaja, 2022). Dalam perjanjian layanan pinjaman uang yang diatur oleh *fintech* berdasarkan POJK No. 77/POJK.01/2016 tentang Layanan Pinjam Meminjam Uang Berbasis Teknologi Informasi (LPMUBT), Pasal 18 POJK mengatur bahwa perjanjian pelaksanaan layanan pinjam meminjam uang berbasis teknologi informasi mencakup dua jenis perjanjian, yaitu perjanjian antara penyelenggara dengan pemberi pinjaman, dan perjanjian antara pemberi pinjaman dengan penerima pinjaman (Hidayat et al., 2022).

Pinjaman *online* memiliki banyak kelebihan yang ditawarkan kepada nasabah salah satunya ialah proses pencairan uang yang cepat dengan syarat yang mudah dan tentunya dapat dilakukan secara daring tanpa harus tatap muka dengan pihak perusahaan namun pinjaman *online* juga memiliki kekurangan yaitu limit kredit yang diberikan kecil atau sedikit dan juga adanya resiko pencurian data nasabah serta tenor pinjaman yang pendek.

2.2 Kredivo

Kredivo merupakan produk kredit instan dari perusahaan PT. FinAccel Finance Indonesia yang diawasi oleh divisi *multifinance* OJK di Indonesia. Kredivo menawarkan kemudahan dalam pembelian dengan sistem "beli sekarang, bayar nanti" dalam waktu 30 hari tanpa dikenakan bunga atau dengan pembayaran cicilan selama 3, 6, atau 12 bulan (dengan bunga sebesar 2,6% per bulan) (Rosiwan & Lasmanah, 2022). Dengan popularitas dan kemudahan yang ditawarkan, Kredivo berhasil meraih banyak pengguna di Indonesia dan menjadi salah satu aplikasi *fintech* yang terkemuka di Indonesia. Namun, seperti halnya layanan keuangan lainnya, Kredivo tetap memerlukan penggunaannya dengan bijak agar tidak mengalami masalah keuangan di masa depan.

2.3 Indodana

Indodana adalah salah satu perusahaan *fintech* atau teknologi keuangan yang didirikan pada tahun 2017. Indodana adalah sebuah aplikasi *fintech* yang menyediakan layanan pinjaman *online* tanpa jaminan di Indonesia. Aplikasi ini diluncurkan pada tahun 2018 oleh PT Artha Dana Teknologi, sebuah perusahaan *fintech* yang berkantor pusat di Jakarta. Indodana ini juga telah terdaftar di OJK berdasarkan Surat OJK S-235/NB.213/2018 sebagai dasar hukum untuk melaksanakan usahanya (Hamsyah & Sulistyowati, 2022).

2.4 AdaKami

AdaKami adalah suatu aplikasi pinjaman *online* yang telah teregistrasi dan diawasi oleh Otoritas Jasa Keuangan (OJK) dan Asosiasi *Fintech* Pendanaan Bersama Indonesia (AFPI). Aplikasi ini telah beroperasi dalam jangka waktu yang cukup lama dan telah menjadi salah satu pilihan utama masyarakat dalam mengunduh aplikasi pinjaman *online* dari Playstore, dengan jumlah unduhan mencapai lebih dari 10 juta (Mochtar & Rusdiana, 2022).

2.5 Akulaku

Akulaku adalah aplikasi *mobile marketplace* yang memungkinkan pengguna untuk mencari toko dan barang yang dijual oleh penjual terdaftar. Aplikasi ini juga menyediakan fasilitas bagi penjual terdaftar untuk menawarkan pembayaran cicilan melalui pembiayaan multi guna untuk pembelian barang yang dijual melalui aplikasi Akulaku (Baiti & Iswandi, 2022). Aplikasi ini diluncurkan pada tahun 2016 oleh PT Akulaku Silvrr Indonesia, sebuah perusahaan teknologi keuangan yang berkantor pusat di Jakarta. Pada awalnya, Akulaku hanya menyediakan layanan *e-commerce* yang memungkinkan para pengguna untuk membeli barang dengan cara mencicil. Namun, seiring berjalannya waktu, Akulaku mulai menyediakan layanan pinjaman *Online* yang memungkinkan para pengguna untuk meminjam uang tanpa jaminan dengan proses yang cepat dan mudah melalui aplikasi.

Dalam beberapa tahun terakhir, Akulaku telah meraih popularitas yang cukup tinggi di Indonesia dan menjadi salah satu aplikasi *fintech* terkemuka di Indonesia. Melalui layanan pinjaman dan *e-commerce* yang mudah dan cepat, Akulaku diharapkan dapat membantu memenuhi kebutuhan finansial dan kebutuhan belanja *online* masyarakat indonesia.

2.6 *Natural Language Processing*

Pemrosesan bahasa alami (*Natural Language Processing*) merupakan cabang dari pembelajaran mesin yang berkaitan dengan pengolahan dan analisis bahasa dan semantik. Dengan menggunakan linguistik komputasional, pembelajaran mesin, dan analisis statistik, algoritma NLP memberikan kemampuan untuk memproses dan memahami teks dengan cara yang lebih mirip dengan manusia dibandingkan model-model lainnya (Hall et al., 2022).

Dalam bidang teknologi informasi, Pemrosesan Bahasa Alami (NLP) merupakan sebuah kemajuan yang memungkinkan untuk membaca umpan balik dalam banyak bahasa dengan minim intervensi manusia. NLP juga mampu melakukan analisis data teks dan mengungkapkan pandangan dan pendapat akhir pengguna terhadap layanan, produk, atau manusia. Dalam hal ini, teknologi NLP

mampu mempercepat proses pengumpulan data serta memberikan informasi yang lebih akurat dan terperinci mengenai persepsi dan pendapat pengguna terhadap suatu layanan atau produk. Hal ini tentu saja memiliki potensi besar dalam meningkatkan kualitas layanan dan produk serta meningkatkan kepuasan pengguna secara keseluruhan (Shaik et al., 2022).

2.7 Analisis Sentimen

Analisis sentimen adalah proses menganalisis sebuah teks atau kata apakah bersifat positif, netral, atau negatif yang kemudian digunakan untuk menentukan nada emosional teks atau kata tersebut. Secara khusus, analisis sentimen melibatkan penggunaan teknik *information retrieval*, NLP, data mining, dan *knowledge management* untuk mengidentifikasi dan mengekstrak informasi subjektif dari volume data tidak terstruktur yang besar (Alsayat, 2022).

Analisis sentimen dapat dilakukan dengan menggunakan metode klasifikasi baik yang *supervised* maupun yang *unsupervised*. Meskipun metode s*upervised* menunjukkan kinerja yang lebih baik dibandingkan dengan metode yang *unsupervised*, namun metode yang *unsupervised* tetap penting karena membutuhkan sedikit data pelatihan berlabel, sementara pengambilan data yang tidak berlabel lebih mudah. Kebanyakan domain, kecuali ulasan film, kekurangan data pelatihan berlabel, sehingga dalam hal ini metode *unsupervised* sangat berguna untuk mengembangkan aplikasi (Vohra & Teraiya, 2013). Analisis sentimen memiliki banyak manfaat, di antaranya:

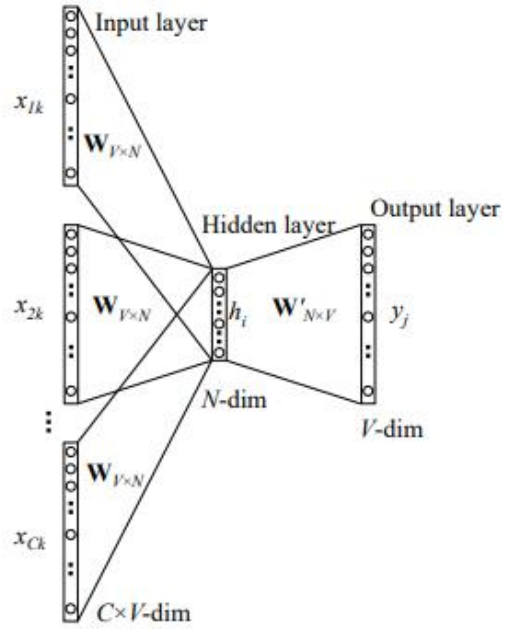
1. Membantu meningkatkan kepuasan pelanggan: Dengan menganalisis sentimen dari umpan balik pelanggan, bisnis dapat memahami sentimen pelanggan terhadap produk dan layanan mereka, serta dapat mengambil tindakan yang sesuai untuk meningkatkan kepuasan pelanggan.
2. Membantu perencanaan pemasaran: Analisis sentimen dapat membantu bisnis memahami opini konsumen tentang produk atau merek mereka, membantu mereka menyesuaikan strategi pemasaran mereka dengan preferensi konsumen.

3. Memantau reputasi merek: Analisis sentimen dapat membantu bisnis memantau apa yang dikatakan orang tentang merek mereka, serta memantau reputasi merek mereka di media sosial dan platform *online* lainnya.
4. Memahami tren pasar: Analisis sentimen dapat membantu bisnis memahami tren pasar dan sentimen konsumen terhadap produk dan merek yang serupa, sehingga dapat mengambil tindakan yang tepat untuk memperkuat posisi mereka di pasar.
5. Membantu manajemen krisis: Analisis sentimen dapat membantu organisasi mengidentifikasi isu-isu yang berkaitan dengan merek mereka dan mengambil tindakan yang cepat dan tepat untuk mengatasi situasi krisis yang mungkin terjadi.
6. Membantu pemantauan opini publik: Analisis sentimen dapat membantu organisasi mengidentifikasi sentimen masyarakat terhadap berbagai topik, sehingga dapat mengambil tindakan yang tepat untuk merespons isu-isu yang sedang ramai diperbincangkan di masyarakat.

2.8 Word2Vec

Word2Vec adalah model NLP (*Natural Language Processing*) yang berfungsi untuk menghasilkan *word embeddings*. Word2Vec pertama kali diperkenalkan oleh Thomas Mikolov dan tim pada tahun 2013 di Google. Tujuan dan manfaat dari Word2Vec adalah untuk mengelompokkan vektor kata yang mirip dalam ruang vektor. Dengan menggunakan data yang cukup, word2vec dapat memprediksi arti kata secara akurat berdasarkan riwayat penggunaannya (Suryati et al., 2023).

Terdapat dua metode yang digunakan pada Word2Vec, yaitu *Continuous Skip-gram* dan *Continuous Bag-of-Words* (CBOW). Perbedaan dari *Continuous Skip-gram* dan *Continuous Bag-of-Words* adalah ada pada cara prediksi kata-katanya. Jika pada *Continuous Skip-gram* memprediksi kata-kata di sekelilingnya dengan diberikan sebuah kata sedangkan pada *Continuous Bag-of-Words* itu memprediksi kata dari konteksnya (Mikolov et al., 2013).

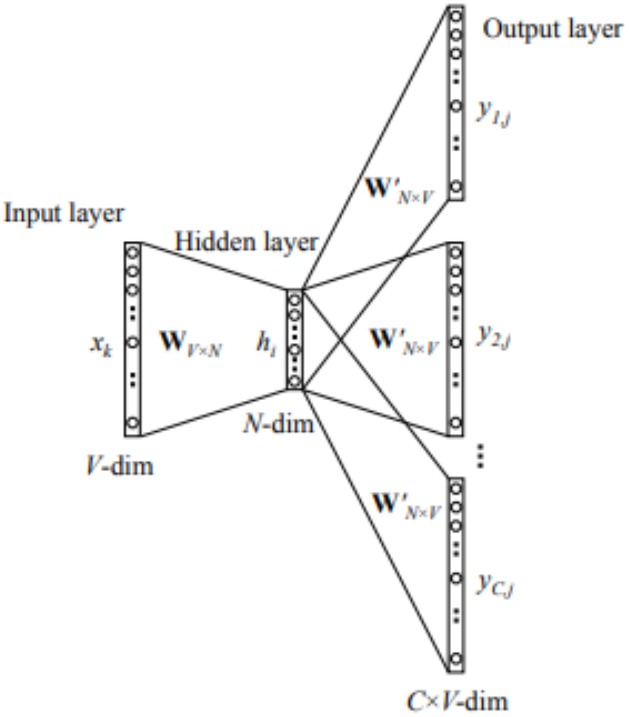


Gambar 2.8.1 Model CBOW (Rong, 2016)

Gambar 2.8.1 menampilkan arsitektur dari model *Continuous Bag-of-Words* (CBOW). *Continuous Bag-of-Words* (CBOW) bertujuan untuk memprediksi kata yang ada di tengah dari kata-kata yang ada di sekitarnya dalam suatu kalimat. Dalam CBOW, setiap kata diwakili oleh vektor *one-hot* dan kemudian diproses melalui sebuah *hidden layer*, yang menghasilkan representasi vektor kata. CBOW cocok digunakan untuk menghasilkan representasi kata yang baik dalam situasi di mana konteksnya tidak terlalu rumit dan hanya dibutuhkan satu representasi kata.

1. *Input Layer*: Pada Word2Vec, input layer mengacu pada representasi vektor kata masukan. Dalam CBOW, input layer menerima vektor-vektor kata konteks yang berada di sekitar kata target, sedangkan dalam Skip-gram, setiap kata dalam konteks menjadi vektor inputnya sendiri.
2. *Hidden Layer*: Hidden layer dalam Word2Vec mengacu pada vektor laten yang terbentuk selama proses pelatihan algoritma. Untuk setiap kata dalam kamus, Word2Vec mempelajari representasi vektor laten yang mencerminkan makna kata tersebut berdasarkan konteksnya. Dalam istilah Word2Vec, *hidden layer* dapat dianggap sebagai representasi vektor "sembunyi" yang berada di antara input dan output.
3. *Output Layer*: Output layer dalam Word2Vec adalah layer yang menghasilkan distribusi probabilitas kata target. Dalam CBOW, output

layer menerima representasi vektor konteks dan berusaha memprediksi kata target. Dalam *Skip-gram*, *output layer* menerima vektor kata target dan berusaha memprediksi konteks sekitarnya.



Gambar 2.8.2 Model *Skip-gram* (Rong, 2016)

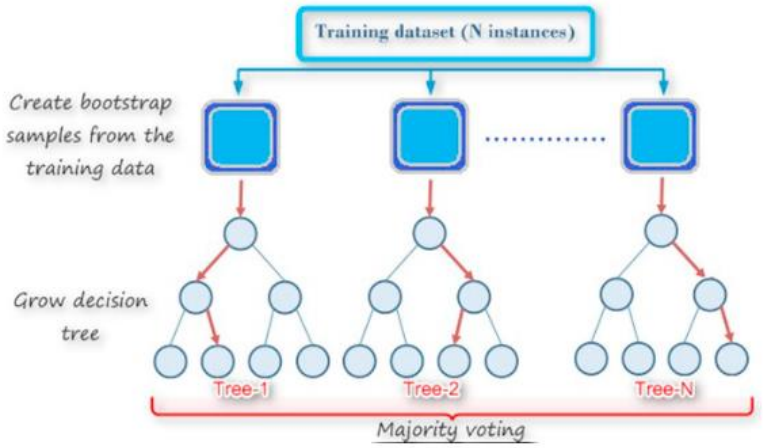
Gambar 2.8.2 menampilkan arsitektur dari model *Skip-gram*. *Continuous Skip-gram* bertujuan untuk memprediksi kata-kata di sekitar suatu kata dalam kalimat, dengan memanfaatkan satu kata sebagai input. *Continuous Skip-gram* memperlakukan kata yang akan diprediksi sebagai input dan menggunakan *hidden layer* untuk menghasilkan representasi vektor untuk setiap kata di sekitarnya.

2.9 Random Forest

Random Forest adalah teknik *machine learning* yang dapat digunakan untuk tugas regresi dan klasifikasi (Rodrigues et al., 2022). Algoritma Random Forest pertama kali diperkenalkan pada tahun 2001 oleh Leo Breiman, seorang profesor di Departemen Statistik University of California, Berkeley. Breiman dan timnya memperkenalkan konsep *ensemble learning*, yang menggabungkan prediksi dari beberapa model *machine learning* yang berbeda untuk meningkatkan akurasi prediksi. Random Forest adalah salah satu jenis *ensemble learning* yang

menggabungkan banyak pohon keputusan (Decision Tree) untuk memperbaiki akurasi prediksi. Konsep Random Forest sendiri terinspirasi oleh algoritma Breiman sebelumnya, yaitu *bagging*, yang juga merupakan teknik *ensemble learning* yang menggabungkan banyak model untuk meningkatkan akurasi.

Random Forest adalah algoritma klasifikasi yang terdiri dari beberapa pohon keputusan atau Decision Tree yang dibangun dengan menggunakan vektor acak (Sandag, 2020). Random Forest adalah algoritma yang termasuk ke dalam teknik *supervised learning* yang pertama kali diperkenalkan oleh Leo Breiman pada tahun 2001 dengan menggabungkan teknik *bootstrap aggregating* dengan *resampling*. Ada beberapa kelebihan dari metode Random Forest yaitu hasil akurasi yang bagus. Dan juga mampu mengatasi *missing value* dan *noise* yang ada pada data, serta algoritma ini cocok untuk mengklasifikasikan data dalam jumlah yang besar.



Gambar 2.9.1 Cara Kerja Algoritma Random Forest (Al Amrani et al., 2018)

Gambar 2.9.1 adalah ilustrasi bagaimana Random Forest bisa menghasilkan sebuah hasil prediksi dengan menggunakan *majority* voting. Dari gambar tersebut dapat dilihat bahwa algoritma Random Forest terdiri dari kombinasi beberapa Decision Trees dimana setiap *tree* bergantung pada nilai random *vector* yang dijadikan sampel secara merata pada semua *tree* yang ada dalam *forest* tersebut. Metode Random Forest memiliki 2 tahapan. Tahap pertama adalah pembentukan *forest* dan tahap kedua adalah *voting* hasil klasifikasi (Willy et al., 2021).

$$forest = \{h(x, \theta_k), k = 1, \dots\} \quad (1)$$

Dimana:

h : Hipotesa atau klasifikasi

x : Input *vector*

θ_k : *Independent and identically distributed random vectors*

Persamaan pertama menyatakan bahwa setiap *forest* terbentuk dari sekumpulan klasifikasi atau hipotesa yang berjumlah k. Input tiap hipotesa adalah x yang kemudian dilakukan *resampling* dengan *random vector* dari x itu sendiri (Willy et al., 2021)

$$C_{rf} = majority\ vote\ \{C_n(x)\} = 1 \quad (2)$$

Dimana:

C_{rf} : *Class* hasil klasifikasi Random Forest

x : Input *vector*

C_n : *Class* prediksi dari *tree* ke-n pada Random Forest

Setelah dilakukan pembuatan *forest,* maka langkah selanjutnya adalah melakukan *voting* untuk klasifikasi dan mengukur performasi dari Random Forest dengan menggunakan persaman kedua (Willy et al., 2021).

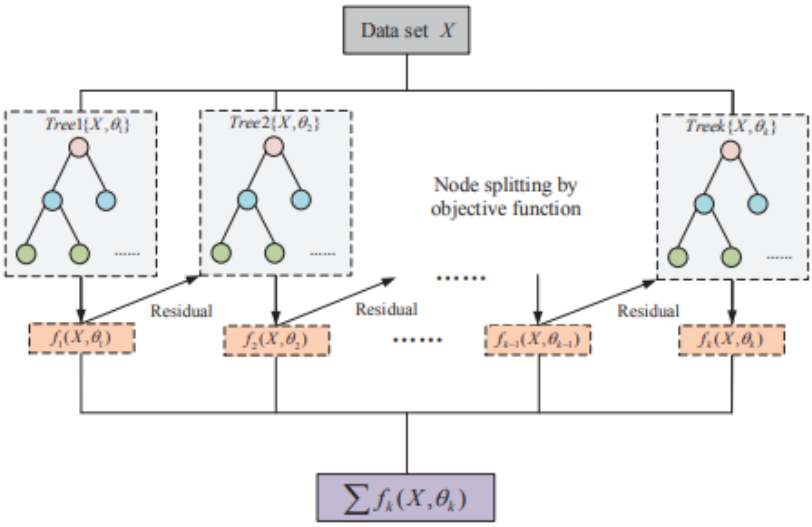
2.10 eXtreme Gradient Boosting (XGBoost)

XGBoost merupakan bagian dari algoritma *boosting* yang terdiri dari kumpulan Decision Tree yang dimana pembentukan setiap *tree* bergantung pada *tree* sebelumnya (Yoris, 2021). XGBoost (eXtreme Gradient Boosting) adalah sebuah algoritma *ensemble learning* untuk *machine learning* yang dikembangkan oleh Tianqi Chen pada tahun 2014. XGBoost adalah implementasi dari algoritma *Gradient Boosting* yang dioptimalkan dengan menggunakan teknik seperti regularisasi dan *pruning*. Sebelumnya, Chen juga telah mengembangkan algoritma *Boosted Trees* yang menjadi dasar dari XGBoost.

Pada tahun 2014, Chen kemudian mengembangkan XGBoost sebagai solusi yang lebih cepat dan lebih efektif dalam menangani masalah *machine learning* yang lebih kompleks. XGBoost mendapatkan popularitasnya dengan kemenangan pada beberapa kompetisi di platform *machine learning* terkemuka seperti Kaggle. Algoritma ini juga mendapatkan dukungan dan kontribusi dari komunitas *open-source*, sehingga terus berkembang hingga saat ini. XGBoost juga menjadi

populer dalam bidang *machine learning* industri dan akademis, digunakan untuk berbagai masalah seperti klasifikasi, regresi, dan ranking (Samih et al., 2023.

XGBoost menggabungkan algoritma pembelajaran berbasis *tree* dan pemecah masalah untuk model linier (Samih et al., 2023). Decision Trees merupakan gabungan dari dua jenis pohon, yaitu *regression tree* dan juga *classification tree*. Jika variabel dependen yang dimiliki bertipe *continue* atau numerik maka Decision Trees menghasilkan *regression tree* sedangkan jika variabel dependen yang dimiliki bertipe kategorik maka Decision Trees menghasilkan *classification tree* (Shafila, 2020).



Gambar 2.10.1 Diagram Skema XGBoost (Guo et al., 2020)

Gambar 2.10.1 menunjukkan alur dari proses komputasi algoritma XGBoost. XGBoost menggunakan model klasifikasi yang dinilai lebih teratur untuk membangun sebuah *regression tree*, XGBoost dapat memberikan kinerja yang lebih optimal dan dinilai mampu untuk mengurangi kompleksitas model sehingga dapat terhindar dari *overfitting*. Nilai akhir dari XGBoost diperoleh dari penjumlahan hasil prediksi dari setiap *regression tree* (Yulianti et al., 2022). Metode klasifikasi XGBoost memerlukan sebuah fungsi objektif yang digunakan untuk menilai tingkat keefektifan model yang diperoleh dari data latih. Ada 2 Karakteristik penting dari sebuah fungsi objektif yaitu *regulation term* dan *training loss*. Fungsi *regulation term* dapat digambarkan seperti pada persamaan (3) berikut ini (Yulianti et al., 2022).

$$obj\ (\theta)\ =\ L\ (\theta)\ +\ \Omega(\theta) \quad (3)$$

Dimana:

L = Fungsi pelatihan yang hilang,

Ω = Nilai regulasi,

θ = Parameter model

Fungsi *training loss* secara umum dapat digambarkan seperti pada persamaan (4) berikut ini.

$$L(\theta)\ =\ \sum_{i=1}^{n} l\ (y_i, \hat{y}_i) \quad (4)$$

Dimana:

y_i = Nilai aktual,

$\hat{y}_i$ = Hasil nilai prediksi dari model,

n = Jumlah iterasi

Untuk melakukan pengukuran *training loss,* dapat digunakan persamaan dari *cross entropy loss* yang dinyatakan dalam persamaan (5) berikut.

$$L(\theta)\ =\ -[y_i log(\hat{y}_i)\ +\ (1\ -\ y_i)\ log(1\ -\ \hat{y}_i)] \quad (5)$$

Kelebihan XGBoost adalah proses komputasi yang cepat, memiliki fleksibilitas yang tinggi, XGBoost juga memiliki fitur regularisasi, dan mampu mengatasi split saat terjadi *negatif loss*.

2.11 Support Vector Machine (SVM)

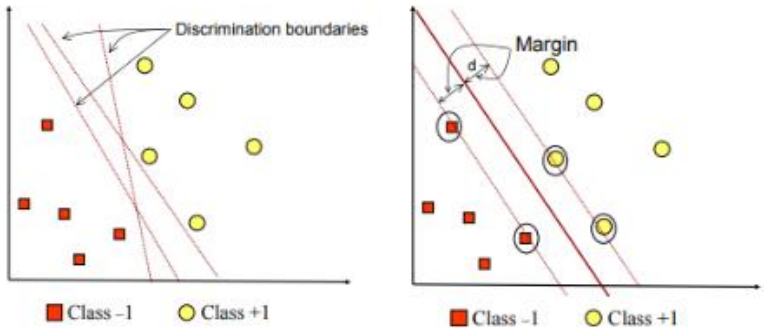
Konsep dasar SVM terletak pada penciptaan *hyperplane* optimal, juga dikenal sebagai batas keputusan atau batas optimal, yang memaksimalkan jarak antara titik data terdekat (*support vector*) dan secara efisien memisahkan kelas-kelas yang berbeda (Adugna et al., 2022). SVM (Support Vector Machine) adalah salah satu algoritma pembelajaran mesin yang populer digunakan dalam klasifikasi dan regresi. Algoritma ini pertama kali diusulkan oleh Vladimir Vapnik dan Alexey Chervonenkis pada tahun 1960-an dan 1970-an saat bekerja di Institut Kibernitika Moskow. Pada awalnya, SVM dikembangkan untuk menyelesaikan masalah klasifikasi dengan dua kelas yang linier terpisah (*linearly separable*), yaitu kasus di mana dua kelompok data dapat dipisahkan dengan garis lurus di antara mereka.

Pada awalnya, SVM hanya digunakan untuk masalah klasifikasi linier, tetapi kemudian dikembangkan untuk menyelesaikan masalah klasifikasi non-linier

melalui penggunaan kernel. Teknik ini memungkinkan SVM untuk melakukan transformasi data mentah ke dalam dimensi yang lebih tinggi sehingga dapat dikelompokkan dengan lebih baik. Pada tahun 1992, Boser, Guyon, dan Vapnik mengusulkan algoritma SVM yang didasarkan pada teori pembelajaran statistik. Kemudian pada tahun 1995, Cortes dan Vapnik memperkenalkan kernel non-linier untuk SVM, yang memungkinkan algoritma ini untuk menangani masalah klasifikasi non-linier dengan sangat baik.

Metode SVM cocok untuk memecahkan masalah yang berhubungan dengan klasifikasi. Awalnya SVM hanya dapat digunakan pada *binary classification* namun kemudian dikembangkan sehingga SVM dapat digunakan untuk menyelesaikan permasalahan klasifikasi yang bersifat *multi-class*. Gambar 2.11.1 menunjukkan sepasang *hyperplane* yang memisahkan dua buah kelas, sedangkan titik kuning dan merah yang terdapat didalam lingkaran hitam disebut dengan *support vector*. Gambar 2.11.1 juga menunjukkan *hyperlane* terbaik karna nilai margin yang dihasilkan besar. Fungsi *hyperplane* yang memisahkan dua kelas dapat dinyatakan dalam persamaan (6) berikut (Cortes & Vapnik, 1995).

$$w_i x_i + b = 0 \quad (6)$$



Gambar 2.11.1 *Hyperplane* (Nugroho, 2008)

Gambar 2.11.1 menjelaskan bagaimana proses pemisahan *hyperplane*. Untuk menghitung margin terbesar dapat dilakukan dengan cara memaksimalkan nilai jarak antara titik terdekat yang ada dengan *hyperplan*e yang dapat dirumuskan sebagai fungsi *Quadratic Programming* (QP) *problem*, yaitu mencari titik minimal dengan menggunakan persamaan (7) dibawah ini (Nugroho, 2008).

$$min\tau(w) = \frac{1}{2}\left|\left|w\right|\right|^2 \quad (7)$$

Dengan syarat memperhatikan *contrain* yang dirumuskan kedalam persamaan berikut.

$$y_i(x_i, w + b) - 1 \geq 0, \forall \mathrm{I} \quad (8)$$

Untuk melakukan optimasi atau menghitung nilai optimal dapat digunakan fungsi *lagrange multiplie* yang dinyatakan kedalam persamaan berikut.

$$L = (w, b, \alpha) = \frac{1}{2}||w|| - \sum_{i=1}^{i} \alpha_i \, (y_i((x_i . w + b) - 1)) \quad (9)$$

α_i dinyatakan sebagai *lagrange multiplie* yang bernilai positif atau nol. Untuk menghitung nilai optimal dari persamaan (9) dapat dilakukan dengan cara meminimalkan nilai L terhadap w dan b. Optimasi juga dapat dilakukan dengan cara memaksimalkan L terhadap α_i dengan memodifikasi persamaan (9) sebagai maksimalisasi problem yang hanya memuat nilai α_i, sehingga menghasilkan persamaan berikut (Nugroho, 2008).

$$\sum_{i=1}^{i} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{i} \alpha_i \alpha_j y_i y_j x_i x_j \quad (10)$$

dimana,

$$\alpha_i \geq 0 \; (i = 1,2,\cdots, l) \; \sum_{i=1}^{i} \alpha_i y_i = 0.$$

Dari perhitungan tersebut didapatkan hasil α_i yang mayoritas bernilai positif. Data yang memiliki kolerasi dengan α_i dan bernilai positif disebut dengan *support vector*. Kelebihan dari SVM adalah cocok digunakan untuk ruang dimensi yang tinggi dan juga hemat memori karena SVM menggunakan *training point* dari fungsi keputusan (*support vector*).

Berikut adalah beberapa kernel pada SVM yang sering digunakan (Srivastava & Bhambhu, 2005).

1. Kernel RBF

$$K(x_i, x_j) = exp \, (-\gamma \parallel x_i^T x_j \parallel^2), \gamma > 0 \quad (11)$$

2. Kernel Poly

$$K(x_i, x_j) = (\gamma(x_i^T x_j) + \mathrm{r})^p \quad (12)$$

3. Kernel Linear

$$K(x_i, x_j) = x_i^T x_j \quad (13)$$

4. Kernel Sigmoid

$$K(x_i, x_j) = tanh(\gamma(x_i^T x_j) + \mathrm{r}) \quad (14)$$

2.12 Particle Swarm Optimization

PSO adalah salah satu teknik optimisasi yang terinspirasi oleh perilaku kelompok dan gerakan koloni dari organisme-organisme dalam alam. Metode ini sering digunakan untuk mencari solusi optimal dalam masalah optimisasi, seperti optimisasi fungsi matematis, pengaturan parameter, penjadwalan tugas, dan pemodelan kecerdasan buatan. Dalam PSO, serangkaian partikel diwakili oleh vektor dalam ruang pencarian multidimensi. Setiap partikel memiliki posisi dan kecepatan saat ini yang menggambarkan solusi yang potensial. Partikel-partikel ini bergerak dalam ruang pencarian untuk mengeksplorasi dan menemukan solusi yang lebih baik dengan berinteraksi satu sama lain.

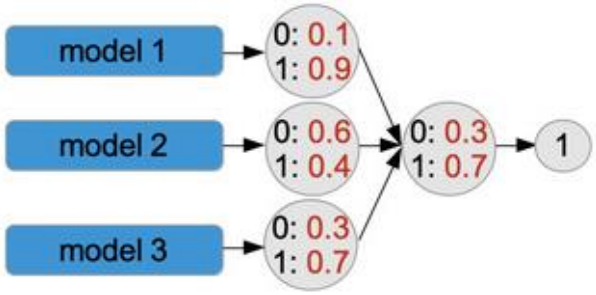
Particle Swarm Optimization (PSO) adalah algoritma metaheuristik yang populer digunakan untuk menyelesaikan masalah optimasi. Algoritma ini terinspirasi oleh perilaku kawanan di alam, seperti kawanan burung atau ikan (Shami et al., 2022). PSO terdiri dari tiga komponen utama: partikel, komponen kognitif dan sosial, dan kecepatan partikel. Setiap partikel mewakili sebuah solusi. Pembelajaran partikel melibatkan dua faktor, yaitu pengalaman partikel dan kombinasi pembelajaran dari seluruh kelompok partikel (Cholissodin & Riyandani, 2016).

Populasi dibentuk secara acak dengan nilai terkecil dan terbesar yang telah ditetapkan sebagai batas bawah dan batas atas. Setiap partikel mencoba menemukan solusi dengan menjelajahi ruang pencarian dan melakukan penyesuaian terhadap posisi terbaik pribadinya (pbest) dan juga penyesuaian terhadap posisi terbaik dari seluruh populasi partikel (gbest) selama proses pencarian. Iterasi dilakukan sejumlah tertentu untuk mencari posisi terbaik setiap partikel, hingga diperoleh posisi yang relatif stabil atau mencapai batas iterasi yang telah ditetapkan. Pada setiap iterasi, performa setiap solusi (posisi partikel) dievaluasi dengan memasukkan solusi tersebut ke dalam fungsi kecocokan (*fitness function*) (Sasongko, 2016).

2.13 *Ensemble Learning*

Ensemble learning adalah sebuah metode yang digunakan untuk menggabungkan beberapa algoritma atau model yang digunakan secara

bersamaan yang bertujuan agar sistem dapat menghasilkan suatu prediksi yang lebih akurat dan tepat daripada hanya menggunakan satu model saja. Salah satu metode *ensemble Classifier* yang populer saat ini adalah *voting Classifier* yang terdiri dari Hard Voting Dan Soft Voting.



Gambar 2.12.1 Soft Voting Classifier (Manconi et al., 2022)

Penelitian ini akan menggunakan metode Soft Voting Classifier. Soft Voting adalah metode *ensemble* yang dinilai lebih kompleks yang pada prosesnya memperhitungkan nilai probabilitas setiap prediksi oleh setiap model klasifikasi atau algoritma (Peppes et al., 2021). Soft Voting Classifier merupakan salah satu teknik *ensemble* learning yang digunakan untuk menggabungkan hasil prediksi beberapa model pembelajaran mesin yang berbeda dalam suatu klasifikasi. Dalam metode klasifikasi Soft Voting, rata-rata probabilitas dari setiap kelas yang ditugaskan oleh setiap model digunakan untuk mengembalikan kelas dengan rata-rata tertinggi dari probabilitas prediksi (Oliveira et al., 2022).

Pada konsep Soft Voting Classifier, setiap model menghasilkan probabilitas atau skor prediksi untuk setiap kelas dan juga setiap model akan menghasilkan sebuah bobot. Kemudian, probabilitas atau skor ini akan dijumlahkan untuk setiap kelas dan diambil rata-ratanya dan dikalikan dengan bobot setiap model untuk mendapatkan nilai akhir probabilitas atau skor untuk setiap kelas. Setelah itu, kelas dengan nilai probabilitas atau skor tertinggi akan diambil sebagai output prediksi akhir. Model-model yang digunakan dapat berbeda jenis atau memiliki konfigurasi parameter yang berbeda. Soft Voting Classifier dapat digunakan pada dataset klasifikasi biner atau multikelas.

Keuntungan dari penggunaan Soft Voting Classifier adalah dapat menghasilkan prediksi yang lebih stabil dan akurat dibandingkan dengan menggunakan satu model saja. Hal ini dikarenakan Soft Voting Classifier

mengambil keuntungan dari kemampuan setiap model dalam memprediksi dan dapat mengkompensasi kekurangan pada setiap model. Selain itu, teknik ini juga dapat meningkatkan *robuster* model terhadap *overfitting*, karena penggunaan beberapa model yang berbeda dapat mengurangi kemungkinan model terlalu menyesuaikan dengan data pelatihan yang spesifik.

2.14 Confusion Matrix

Confusion Matrix adalah sebuah tabel yang digunakan untuk menghitung kinerja model klasifikasi pada satu set data uji yang telah diketehui nilai sebenarnya. Melalui Confusion Matrix kita dapat mengetahui tingkat akurasi, tingkat kesalahan, ketepatan dan nilai penarikan pada suatu model klasifikasi (Saefullah, 2019).

Tabel 2.14.1 Confusion Matrix *Multiclass*

	Predicted			
Actual		Neutral	Negative	Positive
	Neutral	TNeutNeut	ENegNeut	EPosNeut
	Negative	ENeutNeg	TNegNeg	EPosNeg
	Positive	ENeutPos	ENegPos	TPosPos

Tabel 2.13.1 adalah contoh dari Confusion Matrix *Multiclass*, tabel tersebut digunakan untuk menampilkan nilai T*rue Positif* (TP), *False Positif* (FP), *True Negatif* (TN), dan *False Negatif* (FN), dan kemudian dari nilai tersebut dapat diperoleh nilai akurasi, *precission*, dan *recall*.

- TNeutNeut: Variabel ini mewakili jumlah prediksi di mana sentimen netral diklasifikasikan dengan benar [sebagai sentimen Netral]. Ini juga merupakan True Positive untuk kelas neutral.
- ENegNeut: Variabel ini mewakili jumlah prediksi di mana sentimen netral salah diklasifikasikan sebagai sentimen negatif.
- EPosNeut: Variabel ini mewakili jumlah prediksi di mana sentimen netral salah diklasifikasikan sebagai sentimen positif.
- ENeutNet: Variabel ini mewakili jumlah prediksi di mana sentimen negatif salah diklasifikasikan sebagai sentimen netral.

- TNegNeg: Variabel ini mewakili jumlah prediksi di mana sentimen negatif diklasifikasikan dengan benar [sebagai sentimen negatif]. Ini juga merupakan True Positive untuk kelas negatif.
- EPosNeg: Variabel ini mewakili jumlah prediksi di mana sentimen negatif salah diklasifikasikan sebagai sentimen positif.
- ENeutPos: Variabel ini mewakili jumlah prediksi di mana sentimen positif salah diklasifikasikan sebagai sentimen netral.
- ENegPos: Variabel ini mewakili jumlah prediksi di mana sentimen positif salah diklasifikasikan sebagai sentimen negatif.
- TPosPos: Variabel ini mewakili jumlah prediksi di mana sentimen positif diklasifikasikan dengan benar [sebagai sentimen positif]. Ini juga merupakan True Positive untuk kelas positif.

Pada Confusion Matrix, ada 4 istilah yang digunakan untuk merepresentasikan hasil klasifikasi. Keempat istilah tersebut adalah:

1) *True Positve* (TP), merupakan data yang bernilai positif dan hasilnya diprediksi benar. Berikut adalah nilai TP dari masing-masing kelas:
   - TPnetral = TNeutNeut
   - TPnegatif = TNegNeg
   - TPpositif = TPosPos

2) *False Positive* (FP), merupakan data yang bernilai negatif, namun hasilnya diprediksi sebagai data yang bernilai positif. Berikut adalah nilai TN pada masing-masing kelas:

Tabel 2.14.2 Nilai *False Positive*

	Predicted			
Actual		Neutral	Negative	Positive
	Neutral	TNeutNeut	ENegNeut	EPosNeut
	Negative	ENeutNeg	TNegNeg	EPosNeg
	Positive	ENeutPos	ENegPos	TPosPos

Nilai *False Positive* pada kelas netral ditandai dengan warna kuning, pada kelas negative ditandai dengan warna biru dan pada kelas positif ditandai dengan warna merah.

- FPnetral = ENeutNeg + ENeutPos
- FPnegatif = ENegNeut + ENegPos
- FPpositif = EPosNeut + EPosNeg

3) *True Negative* (TN), merupakan data yang bernilai negatif dan hasilnya diprediksi benar. Berikut adalah nilai TN pada masing-masing kelas:

Tabel 2.14.3 Nilai *True Negative* Kelas Netral

	Predicted			
Actual		Neutral	Negative	Positive
	Neutral			
	Negative		TNegNeg	EPosNeg
	Positive		ENegPos	TPosPos

Melalui tabel dapat kita lihat bahwa, *True Negative* untuk kelas netral adalah:

TNnetral = TNegNeg + EPosNeg + ENegPos + TPosPos

Tabel 2.14.4 Nilai *True Negative* Kelas Negatif

	Predicted			
Actual		Neutral	Negative	Positive
	Neutral	TNeutNeut		EPosNeut
	Negative			
	Positive	ENeutPos		TPosPos

Melalui tabel dapat kita lihat bahwa, *True Negative* untuk kelas negatif adalah:

TNnegatif = TNeutNeut + EPosNeut + ENeutPos + TPosPos

Tabel 2.14.5 Nilai *True Negative* Kelas Positif

	Predicted			
Actual		Neutral	Negative	Positive
	Neutral	TNeutNeut	ENegNeut	
	Negative	ENeutNeg	TNegNeg	
	Positive			

Melalui tabel dapat kita lihat bahwa, *True Negative* untuk kelas negatif adalah:

TNpositif = TNeutNeut + ENegNeut +ENeutNeg + TNegNeg

4) *False Negative* (FN), merupakan data yang bernilai positif, namun hasilnya diprediksi sebagai data yang bernilai negatif.

Tabel 2.14.6 Nilai *False Negative*

	Predicted			
Actual		Neutral	Negative	Positive
	Neutral	TNeutNeut	ENegNeut	EPosNeut
	Negative	ENeutNeg	TNegNeg	EPosNeg
	Positive	ENeutPos	ENegPos	TPosPos

Nilai *False Negative* pada kelas netral ditandai dengan warna kuning, pada kelas negatif ditandai dengan warna merah dan pada kelas positif ditandai dengan warna biru.

- FPnetral = ENegNeut + EPosNeut
- FPnegatif = ENeutNeg + EPosNeg
- FPpositi = ENeutPos + ENegPos

Confusion Matrix dapat digunakan untuk menghitung nilai akurasi, presisi, *recall* dan f1-score dari tiap-tiap kelas.

1) Akurasi

Akurasi menggambarkan tingkat keakuratan sistem dalam melakukan proses klasifikasi secara benar. Akurasi dapat dihitung dengan menggunakan persamaan:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (15)$$

2) Presisi

Presisi menggambarkan perbandingan atau rasio antara jumlah prediksi benar bernilai positif dibandingkan dengan keseluruhan hasil yang diprediksi bernilai positif. Nilai presisi dapat dihitung dengan menggunakan persamaan berikut:

$$Presisi = \frac{TP}{TP + FP} \quad (16)$$

3) Recall

Recall menggambarkan seberapa banyak prediksi yang bernilai benar dari semua kelas yang bernilai positif. Nilai *recall* dapat dihitung dengan menggunakan persamaan berikut:

$$Recall = \frac{TP}{TP + FN} \quad (17)$$

4) F1 *Score*

F1 *Score* merupakan perbandingan antara nilai rata-rata presisi dan nilai *recall*. F1 *Score* dapat dinyatakan dengan persamaan berikut:

$$F1\ Score = \frac{2\ x\ Presisi \times Recall}{Presisi + Recall} \quad (18)$$