BAB II
KAJIAN PUSTAKA

2.1 Tinjauan Pustaka

Pada bagian ini diuraikan penelitian sebelumnya yang berkaitan dengan Fintech, fitur, topik model LDA, *imbalance data*, dan klasifikasi CNN.

2.1.1. Penelitian Sebelumnya Terkait Tentang Fintech

Beberapa peneliti melakukan riset tentang pengembangan Fintech menggunakan metode survey di media sosial yang memanfaatkan pengetahuan pelanggan sebagai sumber keunggulan kompetitif bagi perusahaan bisnis (Hidayanti et al., 2018). Pelanggan tidak hanya terlibat dalam transaksi produk, tetapi terlibat dalam proses perusahaan dalam menciptakan produk dan layanan yang sesuai dengan kebutuhan dan keinginan pelanggan. Keterlibatan konsumen dalam proses pengembangan aplikasi memberikan wawasan yang lebih baik tentang kebutuhan konsumen dan meningkatkan efisiensi keseluruhan proses pengembangan aplikasi baru (Zhan et al., 2018). Hasil penelitian pada adopsi Fintech di masa COVID-19 menunjukkan bahwa niat pelanggan untuk menggunakan aplikasi Fintech dipengaruhi oleh persepsinya tentang manfaat, dampak sosial, dan kepercayaan (Al nawayseh, 2020). Kepercayaan pelanggan terhadap Fintech secara signifikan menginterpretasikan hubungan antara risiko yang dirasakan dan niat untuk menggunakan aplikasi Fintech. Hal ini menunjukkan bahwa konsumen akan lebih cenderung melakukan transaksi Fintech ketika ada manfaat yang dirasakan, nilai sosial, dan kepercayaan tinggi, dan pada saat yang sama ketika persepsi risiko rendah. Perkembangan Fintech dari sisi perusahaan telah meningkatkan profitabilitas, inovasi, dan meningkatkan pengendalian risiko bagi bank umum (Y. Wang et al., 2021).

Penelitian Fintech memiliki area riset yang luas dengan dataset yang bervariasi terlihat pada tabel 2.1. Dataset Survei Fintech dan Dataset Artikel Fintech adalah dataset yang paling banyak digunakan dalam penelitian fintech.

Dataset Aktivitas Pinjaman Fintech memiliki potensi besar untuk riset berbasis empiris. Data tingkat transaksi dan aktivitas pengguna sangat sesuai untuk penelitian berbasis *Machine learning* dan Deep Learning. Dataset Teks Komentar Pengguna Fintech memiliki potensi untuk pengembangan model AI yang berfokus pada umpan balik pengguna. Dataset hukum dan transaksi lebih banyak digunakan dalam pendekatan kualitatif dan ekonometrik.

Berikut beberapa penelitian Fintech dengan menggunakan berbagai jenis dataset. Penelitian model Adopsi Fintech menggunakan dataset *survey* untuk menganalisis factor yang mempengaruhi penerimaan dan literasi Fintech digunakan dalam penelitian berbasis Adopsi Teknologi TAM dan UTAUT terlihat pada tabel 2.2 Studi yang menggunakan metode ini antar lain (Al-Omoush & Alsmadi, 2024; Alnawayseh, 2020; Chawla & Joshi, 2020; Ha & Nguyen, 2024; Krah et al., 2024; Shaikh & Amin, 2024; A. Sharma et al., 2024). Penelitian menggunakan Analisis Bibliometrik menggunakan dataset tren penelitian Fintech untuk menyelidiki evolusi FinTech dalam penelitian ilmiah terlihat pada tabel 2.3, metode ini digunakan oleh (Abad-Segura et al., 2020; AlQudah et al., 2024; Dissanayake et al., 2023; Garad et al., 2025; Sahabuddin et al., 2023). Penelitian model Studi empiris Fintech menggunakan regresi panel, SEM (Structural Equation Modeling) dan OLS menggunakan dataset aktivitas pinjaman Fintech, matrik kinerja Fintech untuk menganalisa pengaruh sosial, inovasi, efisiensi, dan kinerja FinTech terlihat pada tabel 2.4 penelitian ini dilakukan oleh (Ha & Nguyen, 2024; J. Li et al., 2025; Y. Li et al., 2025; Liao et al., 2025; P. Wang & Yang, 2024; Xu et al., 2025) Penelitian model *Machine learning* dan Deep Learning pada Fintech menggunakan dataset transaksi fintech dan intellectual property office menggunakan berbagai algoritma untuk mendeteksi penipuan, peramalan, prediksi dan pengambilan keputusan FinTech terlihat pada tabel 2.5 (T. H. Chen & Chang, 2021; Gambacorta et al., 2024; Shehadeh, 2025; Stojanovi et al., 2021; Upreti et al., 2023; H. Zhang et al., 2024). Penelitian Analisis Kualitatif pada Fintech menggunakan dokumen hukum, seperti wawancara, survei, dan tinjauan literatur untuk memahami faktor non-kuantitatif dalam FinTech yaitu menganalisis tantangan regulasi FinTech, pengaruh FinTech pada lembaga keuangan mikro, dan menjelaskan peluang dan tantangan dalam

pengembangan FinTech terlihat pada tabel 2.6. penelitian kualitatif dilakukan oleh (Ally, 2025; Garad et al., 2025; Offiong et al., 2024). Penelitian analisis ekonometrika pada Fintech menggunakan dataset tentang perkembangan FinTech dan indikator stabilitas keuangan untuk menganalisis dampak FinTech pada ketidakstabilan keuangan, pertumbuhan berkelanjutan, dan pengembangan dan implementasi berbagai solusi FinTech terlihat pada tabel 2.7. penelitian ini dilakukan oleh (Guermazi, 2025; Ioannou et al., 2024; Lin & Song, 2025)

Dari penelitian sebelumnya terlihat bahwa penelitian menggunakan dataset Fintech memiliki berbagai jenis data dan pendekatan metode penelitian yang bervariasi. Dataset teks komentar pengguna Fintech untuk pengembangan model berbasis *Machine learning* merupakan area riset yang masih belum banyak dieksplorasi. Pada penelitian ini diusulkan dengan menggunakan pengolahan data teks komentar pengguna Fintech untuk menggali fitur layanan dengan menggunakan *machine learning* LDA dan CNN.

SEKOLAH PASCASARJANA

Tabel 2.1 Metode Penelitian dan Jenis Dataset FIntech

Metode Penelitian / Dataset	Adopsi Teknologi TAM UTAUT	Analisis Bibliometrik	Studi Empiris	*Machine learning* dan Deep Learning	Analisis Kualitatif	Analisis Ekonometrika
Dataset Survey Fintech	(Al-Omoush & Alsmadi, 2024; Alnawayseh, 2020; Chawla & Joshi, 2020; Ha & Nguyen, 2024; Krah et al., 2024; Shaikh & Amin, 2024; A. Sharma et al., 2024)					
Dataset Artikel Fintech		(Abad-Segura et al., 2020; AlQudah et al., 2024; Dissanayake et al., 2023; Garad et al., 2025; Sahabuddin et al., 2023)				
Dataset Aktivitas Pinjaman Fintech			(Ha & Nguyen, 2024; J. Li et al., 2025; Y. Li et al., 2025; Liao et al., 2025; P. Wang & Yang, 2024; Xu et al., 2025)			
Data Transaksi Fintech dan Intellectual Property Office				(T. H. Chen & Chang, 2021; Gambacorta et al., 2024; Shehadeh, 2025; Stojanovi et al., 2021; Upreti et al., 2023; H. Zhang et al., 2024)		
Dokumen Hukum, dan Analisis Data Fintech					(Ally, 2025; Garad et al., 2025; Offiong et al., 2024)	
Pengembangan dan implementasi berbagai solusi FinTech						(Guermazi, 2025; Ioannou et al., 2024; Lin & Song, 2025)
Dataset teks Komentar Pengguna Fintech				Diusulkan untuk diteliti LDA-CNN		

SEKOLAH PASCASARJANA

Tabel 2.2 SLR Adopsi teknologi Fintech

No	Judul	Tujuan Penelitian	Dataset	Metode Penelitian	Temuan Utama
1	Consumers' innovativeness and acceptance towards use of financial technology in Pakistan: extension of the UTAUT model (Shaikh & Amin, 2024)	Menganalisis faktor yang mempengaruhi penerimaan nasabah layanan FinTech dan Menguji pengaruh Consumer Innovatiness dalam memoderasi penerimaan FinTech	Jumlah Sampel: 216 responden (nasabah bank di Pakistan). Metode Pengambilan Sampel: Judgemental Sampling. Periode: 1 Desember 2021 – 2 Februari 2022.	. Model UTAUT dengan variabel	Model UTAUT yang dimodifikasi dengan variabel Consumer Innovativeness memperkuat prediktabilitas penerimaan FinTech.
2	Unlocking financial access in a developing country amidst COVID-19: the impacts of financial literacy and fintech (Ha & Nguyen, 2024)	Mengetahui dampak teknologi keuangan dan literasi keuangan terhadap pengembangan inklusi keuangan di kalangan generasi muda	kuesioner semi-terstruktur menggunakan alat survei daring Google. 1300 peserta	Regresi logistik untuk menyelidiki dampak literasi keuangan,	kesenjangan dalam akses ke layanan keuangan antara daerah perkotaan dan pedesaani. Teknologi finansial dan literasi keuangan memainkan peran penting dalam memfasilitasi tingkat pengembangan inklusi keuangan
3	Financial Technology Adoption Among Small and Medium Enterprises (SMEs) in Ghana; (Krah et al., 2024)	Menyelidiki faktor-faktor yang mempengaruhi adopsi FinTech di kalangan UKM di Ghana.	Data survei dari UKM di Ghana.	Penerapan Model Penerimaan Teknologi (TAM).	Mengidentifikasi faktor kunci yang mempengaruhi adopsi FinTech, memberikan wawasan untuk pembuat kebijakan dan praktisi.
4	Determinants of fintech adoption in agrarian economy: Study of UTAUT extension model in reference to developing economies (A. Sharma et al., 2024)	memperluas model UTAUT2 dalam konteks ekonomi agraria+ Behavioral (BIAFPS) sebagai mediator.	362 petani dari 6 distrik di negara bagian Uttar Pradesh, India.	Berdasarkan UTAUT2 Menambahkan variabel Behavioral Intention to Adopt FinTech Products and Services (BIAFPS) sebagai mediator untuk memperkuat prediktabilitas model.	Social Influence (SI), Performance Expectancy (PE), dan Convenience (C) memiliki pengaruh positif dan signifikan terhadap BIAFPS dan AAFPS. Adopsi FinTech mampu meningkatkan efisiensi dan profitabilitas.

SEKOLAH PASCASARJANA

Tabel 2.2 SLR Adopsi teknologi Fintech lanjutan

No	Judul	Tujuan Penelitian	Dataset	Metode Penelitian	Temuan Utama
5	Role of Mediator in Examining the Influence of Antecedents of Mobile Wallet Adoption on Attitude and Intention (Chawla & Joshi, 2020)	Menjelaskan faktor-faktor utama yang memengaruhi adopsi mobile wallet di India. Menguji peran mediasi dari faktor-faktor seperti perceived usefulness, trust, dan attitude dalam memperkuat hubungan antara faktor eksternal dengan niat adopsi.	Jumlah sampel: 744 responden 66,3% laki-laki dan 33,7% perempua	Berdasarkan TAM dan UTAUT	Penelitian ini memperluas model TAM dan UTAUT dengan menambahkan peran mediasi dari trust, perceived usefulness, dan attitude.
6	FinTech in COVID-19 and Beyond: What Factors Are Affecting Customers' Choice of FinTech Applications? (Alnawayseh, 2020)	Mengidentifikasi dan menganalisis faktor-faktor yang mempengaruhi niat pengguna di Yordania untuk mengadopsi aplikasi FinTech selama pandemi COVID-19.	Data dikumpulkan melalui survei daring di Yordania pada Mei 2020 selama periode lockdown akibat COVID-19. Total respons = 500	Menggunakan model modifikasi dari UTAUT yang dipadukan dengan Extended Valence Framework untuk mengukur pengaruh	Persepsi manfaat dan pengaruh sosial adalah faktor utama yang meningkatkan niat pengguna dalam menggunakan aplikasi FinTech di Yordania selama COVID-19. Persepsi risiko tidak mempengaruhi niat penggunaan secara langsung, tetapi mempengaruhi kepercayaan pengguna terhadap aplikasi FinTech.

SEKOLAH PASCASARJANA

Tabel 2.3 SLR Analisis Bibliometrik

No	Judul	Tujuan Penelitian	Dataset yang Digunakan	Metode Penelitian	Temuan Utama
1	Financial Technology: Review of Trends, Approaches, and Management (Abad-Segura et al., 2020)	Menganalisis tren penelitian global dalam teknologi keuangan dari 1975 hingga 2019.	Sampel 2012 artikel dari jurnal ilmiah yang dipilih dari database Scopus Elsevier.	Analisis bibliometrik.	Mengidentifikasi kontribusi utama (penulis, institusi, negara) dan tema utama dalam penelitian teknologi keuangan.
2	A Bibliometric Analysis of Financial Technology: Unveiling the Research Landscape (Dissanayake et al., 2023)	Menyediakan analisis bibliometrik komprehensif tentang penelitian FinTech, menyoroti tren dan faktor yang berpengaruh.	1373 publikasi dari database Scopus mencakup bisnis, manajemen, akuntansi, ekonomi, keuangan, dan ilmu sosial.	Analisis bibliometrik menggunakan perangkat lunak biblioshiny, Hukum Lotka, dan Hukum Bradford.	Mengungkap peningkatan output penelitian, ketertarikan global, kontributor utama, dan tema yang berkembang seperti teknologi blockchain dan pembayaran digital.
3	The Evolution of FinTech in Scientific Research: A Bibliometric Analysis (Sahabuddin et al., 2023)	Menyelidiki evolusi penelitian FinTech dari waktu ke waktu menggunakan analisis bibliometrik.	1359 publikasi terkait FinTech dari 2010 hingga 2021, dikumpulkan dari database Scopus.	Analisis bibliometrik mencakup analisis tema, analisis kemunculan bersama, dan analisis garis waktu pada kata kunci penulis.	Menyoroti tren publikasi tahunan yang meningkat, pergeseran fokus pada inklusi keuangan, dominasi penulis dari AS, dan meningkatnya kolaborasi internasional.
4	Financial Technology's Role in Advancing Social Responsibility: A Bibliometric Review of Research Progress and Future Opportunities (AlQudah et al., 2024)	Mengidentifikasi dan mengkaji faktor kunci dalam FinTech yang mempengaruhi tanggung jawab sosial.	Kompilasi publikasi terkait FinTech dan tanggung jawab sosial dari berbagai basis data akademik.	Analisis bibliometrik.	Mengidentifikasi faktor utama di mana FinTech berkontribusi terhadap tanggung jawab sosial dan memberikan wawasan untuk pengembangan kebijakan di masa depan.

SEKOLAH PASCASARJANA

Tabel 2.4 SLR Studi Empiris

No	Judul	Tujuan Penelitian	Dataset	Metode Penelitian	Temuan Utama
1	The Impact of FinTech Development on Liquidity Mismatch in Banks Emerald (Liao et al., 2025)	Menyelidiki bagaimana pengembangan FinTech mempengaruhi ketidakcocokan likuiditas di bank.	Data tentang aktivitas pemberian pinjaman bank, termasuk metode investigasi kredit inovatif seperti big data analytics.	Analisis empiris yang menguji hubungan antara kemajuan FinTech dan pengelolaan likuiditas di lembaga perbankan.	FinTech mendorong pengembangan teknologi pinjaman melalui big data, sehingga memperluas akses kredit dan mempengaruhi posisi likuiditas di bank.
2	FinTech, Financial Development, and Banking Efficiency (Y. Li et al., 2025)	Menyelidiki dampak FinTech terhadap efisiensi perbankan di Tiongkok, dengan mempertimbangkan perbedaan dalam perkembangan keuangan.	Data tentang perkembangan FinTech dan matrik kinerja perbankan di Tiongkok.	Analisis empiris untuk menguji hubungan antara kemajuan FinTech dan efisiensi perbankan.	FinTech secara positif mempengaruhi efisiensi perbankan, dengan variasi berdasarkan tingkat perkembangan keuangan.
3	Social trust and household financial decision-making: An empirical study based on the usage of household financial technology (P. Wang & Yang, 2024)	Menganalisis pengaruh social trust terhadap keputusan keuangan rumah tangga dalam konteks penggunaan teknologi keuangan.	China Household Finance Survey (CHFS) - Periode data: 2011, 2013, 2015, dan 2017 - Jumlah Observasi: 16.631 data panel dari 29 provinsi	Model regresi panel: Instrumental Variable (IV) Regression, Robustness Test, Mechanism Test, Heterogeneity Test	Social trust meningkatkan keputusan keuangan rumah tangga dalam bentuk peningkatan pengeluaran belanja online Pengetahuan keuangan memediasi pengaruh social trust terhadap keputusan keuangan

SEKOLAH PASCASARJANA

Tabel 2.4 SLR Studi Empiris lanjutan

No	Judul	Tujuan Penelitian	Dataset	Metode Penelitian	Temuan Utama
4	How Does Financial Technology Innovation Impact Business Performance? (J. Li et al., 2025)	Menguji dampak inovasi FinTech pada kinerja bisnis.	Data tentang inovasi FinTech dan matrik kinerja bisnis.	Analisis empiris untuk menilai hubungan antara inovasi FinTech dan hasil bisnis.	Inovasi FinTech secara signifikan berkontribusi pada peningkatan kinerja bisnis dengan meningkatkan transparansi informasi dan mengurangi kendala pembiayaan.
5	Does FinTech Foster the Development of Green Finance? (Xu et al., 2025)	Menguji dampak FinTech pada pengembangan keuangan hijau di Tiongkok.	Data panel dari 30 provinsi di Tiongkok (2011–2021).	Analisis empiris menggunakan model regresi.	Perkembangan FinTech secara positif mempengaruhi keuangan hijau, meningkatkan pertumbuhan ekonomi yang berkelanjutan.
6	Unlocking financial access in a developing country amidst COVID-19: the impacts of financial literacy and fintech (Ha & Nguyen, 2024)	Mengetahui dampak teknologi keuangan dan literasi keuangan terhadap pengembangan inklusi keuangan di kalangan generasi muda	kuesioner semi-terstruktur menggunakan alat survei daring Google. 1300 peserta	regresi logistik untuk menyelidiki dampak literasi keuangan,	kesenjangan dalam akses ke layanan keuangan antara daerah perkotaan dan pedesaani. Teknologi finansial dan literasi keuangan memainkan peran penting dalam memfasilitasi tingkat pengembangan inklusi keuangan
7	Unraveling the Nexus: Intellectual Capital, FinTech Innovation, Competitive Agility, and Financial Inclusion (Al-Omoush & Alsmadi, 2024)	Mengeksplorasi hubungan antara modal intelektual, inovasi FinTech, ketangkasan kompetitif, dan inklusi keuangan.	Data dari lembaga keuangan tentang matrik modal intelektual dan indeks inovasi FinTech.	Analisis empiris menggunakan pemodelan persamaan struktural (SEM).	Modal intelektual dan inovasi FinTech secara signifikan meningkatkan ketangkasan kompetitif dan inklusi keuangan.

SEKOLAH PASCASARJANA

Tabel 2.5 SLR Machine learning dan Deep Learning

No	Judul	Tujuan Penelitian	Dataset yang Digunakan	Metode Penelitian	Temuan Utama
1	Follow the Trail: *Machine learning* for Fraud Detection in FinTech (Stojanovi et al., 2021)	Memberikan gambaran tentang metode pembelajaran mesin untuk deteksi penipuan dalam domain FinTech.	Tidak disebutkan secara spesifik; berfokus pada metodologi.	Tinjauan tentang metode pembelajaran mesin, termasuk teknik deteksi anomali, algoritma pembelajaran, model statistik, dan jaringan saraf buatan.	Membahas penerapan metode pembelajaran mesin dalam deteksi penipuan, menyoroti tantangan seperti kesalahan positif dan kebutuhan akan teknik pencegahan penipuan yang efektif.
2	Using *Machine learning* to Evaluate the Influence of FinTech Patents: The Case of Taiwan's Financial Industry (T. H. Chen & Chang, 2021)	Menganalisis pengaruh paten FinTech terhadap kinerja industri keuangan di Taiwan. Mengevaluasi peran paten FinTech	Sumber Data: Intellectual Property Journal Database (TEJ). - Periode: 2008 – 2016. - Jumlah Sampel: 333	Regression Analysis: Model regresi untuk mengukur pengaruh paten FinTech pada kinerja keuangan (ROA dan ROE). Dan MLP	Model *Machine learning* (ML) lebih akurat dibandingkan regresi tradisional; *Machine learning* lebih efektif dalam memprediksi pengaruh paten FinTech dibandingkan metode regresi.
3	Exploring Cutting-Edge Algorithms in FinTech: Leveraging *Machine learning* for Financial Forecasting (Shehadeh, 2025)	Mengeksplorasi penerapan algoritma pembelajaran mesin tingkat lanjut dalam peramalan keuangan di sektor FinTech.	Dataset keuangan besar yang mencakup berbagai indikator keuangan dan variabel pasar.	Penerapan jaringan saraf buatan (ANN) dan teknik pembelajaran mesin lainnya untuk memodelkan hubungan kompleks dan non-linear dalam data keuangan.	ANN secara efektif menangani hubungan kompleks dalam dataset besar, meningkatkan akurasi peramalan dan pengambilan keputusan keuangan dalam perusahaan FinTech.

SEKOLAH PASCASARJANA

Tabel 2.5 SLR Machine learning dan Deep Learning lanjutan

No	Judul	Tujuan Penelitian	Dataset yang Digunakan	Metode Penelitian	Temuan Utama
4	How do *machine learning* and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm (Gambacorta et al., 2024)	Membandingkan kinerja model penilaian kredit berbasis *machine learning* dengan model tradisional dalam memprediksi risiko kredit dan tingkat gagal bayar.	Data tingkat transaksi dari perusahaan fintech terkemuka di China. Periode: Mei 2017-September 2017. Jumlah Observasi: 310.919 data pinjaman.	Model Tobit untuk memprediksi tingkat kerugian pinjaman. Model Logit untuk memprediksi kemungkinan gagal bayar.	Model *machine learning* berbasis data non-tradisional lebih akurat dalam memprediksi risiko kredit *Machine learning* meningkatkan kemampuan prediksi model sebesar 5,3% setelah kejutan ekonomi.
5	Enhanced Algorithmic Modelling and Architecture in Deep Reinforcement Learning Based on Wireless Communication Fintech Technology (Upreti et al., 2023)	Mengembangkan protokol baru untuk pengiriman data yang aman dengan menggunakan fuzzy rule-based secure transmission dan DRNN.	Bank Transaction Data: Data terdiri dari 594.643 transaksi (587.443 transaksi normal dan 7.200 transaksi penipuan).	Fuzzy Rule-Based Secure Transmission: Deep Reinforcement Neural Network (DRNN)	Protokol yang diusulkan lebih unggul dibandingkan metode SVM dan RF. DRNN menunjukkan kinerja tinggi Model berbasis fuzzy dan DRNN meningkatkan efisiensi proses pengambilan keputusan dalam pengiriman data keuangan yang kompleks.
6	A *Machine learning* and Quantile Analysis of FinTech and Resource Efficiency in Achieving Sustainable Development in OECD Countries (H. Zhang et al., 2024)	Menganalisis dampak investasi FinTech dan efisiensi sumber daya terhadap pembangunan berkelanjutan di negara-negara OECD.	PitchBook, OECD Data, World Development Indicators (WDI), United Nations Environment Program. Periode: 2010 – 2019	Method of Moments Quantile Regression (MMQREG) Kernel Regularized Least Squares (KRLS)	FinTech berpengaruh positif pada pembangunan berkelanjutan di kuantil tinggi. Efisiensi sumber daya (RESEFF) memiliki pengaruh positif. Kualitas Institusi (IQ) memiliki dampak positif dan signifikan. Teknologi ramah lingkungan (EVTEC) berpengaruh positif.
7	Model Analytic in Fintech User Comment Features Using LDA-CNN on Imbalanced Data.	Mengembangkan model analitik gabungan algoritma LDA-ROS-NCL-CNN untuk menganalisis komentar pengguna Fintech	Google Play Komentar Pengguna Fintech di Indonesia	LDA-ROS-NCL-CNN	Mengidentifikasi topik dan fitur utama. integrasi LDA-CNN untuk analisis topik dan klasifikasi, pendekatan baru dalam konteks pengembangan fitur fintech

SEKOLAH PASCASARJANA

Tabel 2.6 SLR Analisis Kualitatif

No	Judul	Tujuan Penelitian	Dataset yang Digunakan	Metode Penelitian	Temuan Utama
1.	Artificial Intelligence (AI) and Financial Technology (FinTech) in Tanzania: Legal and Regulatory Issues (Ally, 2025)	Menganalisis tantangan hukum dan regulasi yang terkait dengan adopsi AI dan FinTech di Tanzania.	Dokumen hukum, kerangka regulasi, dan studi kasus dari Tanzania.	Analisis kualitatif terhadap teks hukum dan regulasi.	Mengidentifikasi kesenjangan dalam regulasi yang ada dan mengusulkan kerangka kerja untuk mengakomodasi teknologi keuangan baru di sektor keuangan Tanzania.
2.	FinTech as a Digital Innovation in Microfinance Companies (Offiong et al., 2024)	Melakukan tinjauan literatur tentang FinTech di lembaga keuangan mikro untuk meningkatkan efisiensi operasional dan pengalaman pelanggan.	Literatur dari berbagai studi tentang aplikasi FinTech di lembaga keuangan mikro.	Tinjauan literatur sistematis.	Inovasi FinTech secara signifikan meningkatkan efisiensi operasional dan kepuasan pelanggan di lembaga keuangan mikro.
3.	An Overview of FinTech: Opportunities, Challenges, and Potential Development (Garad et al., 2025)	Memberikan gambaran umum tentang lanskap FinTech, termasuk peluang, tantangan, dan prospek pengembangan masa depan.	Analisis data keuangan yang ada dan integrasi teknologi seperti AI dan IoT.	Tinjauan literatur dan penilaian kualitatif terhadap integrasi teknologi dalam layanan keuangan.	Kombinasi teknologi AI dan IoT dalam layanan keuangan modern dapat mengekstraksi lebih banyak data yang bernilai, meningkatkan layanan pelanggan dan efisiensi operasional.

SEKOLAH PASCASARJANA

Tabel 2.7 SLR Analisis Ekonometrika

No	Judul	Tujuan Penelitian	Dataset	Metode Penelitian	Temuan Utama
1.	FinTech and Financial Instability: Is This Time Different? (Ioannou et al., 2024)	Mengkaji perkembangan FinTech sebagai sektor yang menerapkan inovasi digital dan implikasinya terhadap ketidakstabilan keuangan.	Data tentang perkembangan FinTech dan indikator stabilitas keuangan.	Studi analitik yang mengeksplorasi hubungan antara pertumbuhan FinTech dan ketidakstabilan keuangan.	Mengidentifikasi risiko dan tantangan regulasi yang muncul akibat pertumbuhan pesat FinTech.
2.	Impact of Financial Technology (FinTech) on Sustainable Growth (Guermazi, 2025)	Menyelidiki dampak FinTech pada pertumbuhan berkelanjutan, dengan fokus pada evolusi solusi FinTech dan faktor-faktor yang mempengaruhinya.	Data tentang pengembangan dan implementasi berbagai solusi FinTech selama beberapa tahun.	Studi analitik yang memeriksa korelasi antara kemajuan FinTech dan pertumbuhan ekonomi yang berkelanjutan.	Pengembangan FinTech secara positif mempengaruhi pertumbuhan berkelanjutan dengan meningkatkan inklusi keuangan dan efisiensi operasional.
3.	Financial Technology and Corporate Idiosyncratic Volatility (Lin & Song, 2025)	Menyelidiki pengaruh aplikasi FinTech pada volatilitas idiosinkratik perusahaan.	Data tentang penggunaan aplikasi FinTech dan matrik volatilitas perusahaan.	Analisis regresi OLS (Ordinary Least Squares).	Aplikasi FinTech secara signifikan mempengaruhi pengurangan volatilitas idiosinkratik perusahaan.

SEKOLAH PASCASARJANA

Fitur merupakan seperangkat fungsionalitas sistem yang kohesif subset dari persyaratan sistem, subset implementasi sistem. Subset dari persyaratan sistem yaitu persyaratan yang mewakili semua karakteristik perilaku penting dari suatu sistem. (Reid Turner et al., 1999). Fitur adalah elemen utama yang terlihat oleh pengguna dan memberikan fungsi atau karakteristik yang membedakan satu aplikasi dari aplikasi lainnya (Igler, 2013). Fitur aplikasi mencakup elemen-elemen utama yang tersedia di dalam aplikasi untuk mendukung pengelolaan layanan (Tao et al., 2023).

Penelitian fitur aplikasi terlihat pada tabel 2.8 antara lain rekomendasi fitur dari UI dan deskripsi aplikasi dari penelitian (Lo, 2019) dan (Jiang et al., 2019) menunjukkan bahwa kombinasi metode berbasis UI similarity dan deskripsi aplikasi sangat efektif dalam meningkatkan strategi pengembangan fitur. Algoritma Genetika (GA) dan Model Topik (LDA) meningkatkan akurasi dalam mencocokkan fitur dengan aplikasi serupa. Pembaruan Fitur Berdasarkan Produk Serupa yang dilakukan oleh Liu et al. (2021) menunjukkan kombinasi metode LDA (untuk identifikasi aplikasi serupa) dan SVM (untuk klasifikasi kalimat dalam teks rilis) efektif dalam memperbarui fitur. Penggunaan teks rilis dan umpan balik pengguna mempercepat proses pembaruan fitur. Pemanfaatan Ulasan Pengguna untuk Pengembangan Fitur dilakukan oleh Palomba et al. (2018) menunjukkan bahwa penggunaan model CRISTAL meningkatkan kemampuan pengembang untuk mengambil keputusan berdasarkan ulasan pengguna. Analisis Domain dan Keterlacakan Fitur dilakukan oleh Hariri et al. (2013) dan Reid Turner (1999) menunjukkan bahwa kombinasi Incremental Diffusive Clustering (IDC) dan Association Rule Mining meningkatkan presisi dalam pengelompokan fitur. Penelitian terdahulu tentang fitur mengusulkan peningkatan kemampuan peningkatan kemampuan dan efektivitas Fintech, namun masih bersifat umum dengan membandingkan deskripsi aplikasi aplikasi satu dengan lainnya, belum ada yang memanfaatkan fitur untuk layanan Fintech, maka penelitian ini mengusulkan penggalian fitur aplikasi Fintech, berdasarkan komentar pengguna.

SEKOLAH PASCASARJANA

Tabel 2.8 SLR Pengembangan Fitur

No	Judul	Tujuan Penelitian	Dataset	Metode Penelitian	Temuan Utama
1	Attentive Aspect Modeling for Review-Aware Recommendation (Chua, 2019)	Mengatasi dua masalah utama dalam sistem rekomendasi berbasis aspek: Sparsity aspek, Variasi perhatian pengguna	Amazon Product Dataset (subkategori "five-core")	Attentive Aspect-based Recommendation Model (AARM) User-Level Attention Module, Bayesian Personalized Ranking (BPR)	AARM mengungguli DeepCoNN, JRL-Review) dan multi-modal (JRL, eJRL). Peran Penting dari Perhatian Pengguna yang Bervariasi
2	Feature Evaluation for Mobile Applications: A Design Science Approach Based on Evolutionary Software Prototypes(Igler, 2013)	Mengidentifikasi fitur utama dalam pengembangan aplikasi mobile dalam konteks bisnis tertentu. Mengembangkan model proses Identifikasi fitur antarmuka pengguna (UI)	Proyek SMAT (Success Factors of Mobile Application Design for Public Transportation)	Design Science Research (DSR) Relevance Cycle, Design Cycle, Rigor Cycle	Model proses berbasis fitur memungkinkan pengembangan prototipe aplikasi yang dapat dievaluasi dan diperbaiki secara bertahap; Sinkronisasi model fitur dengan prototipe
3	Crowdsourcing User Reviews to Support the Evolution of Mobile Apps (Palomba et al., 2018)	Mengukur sejauh mana pengembang aplikasi mengambil umpan balik dari ulasan pengguna dalam proses pengembangan dan perbaikan aplikasi.	100 aplikasi open-source Android dari Google Play Store	Pengembangan CRISTAL (Crowdsourcing RevIews to SupporT App evoLution)	Sebagian besar ulasan pengguna yang informatif dipertimbangkan: 49% ulasan informatif diterapkan pada rilis aplikasi berikutnya. 75% pengembang mengakui sering mempertimbangkan ulasan pengguna dalam pengembangan aplikasi.
4	Supporting Domain Analysis through Mining and Recommending Features from Online Product Listings (Hariri et al., 2013)	Mengembangkan sistem rekomendasi fitur untuk mendukung proses analisis domain dalam pengembangan perangkat lunak	Dataset utama berasal dari situs web Softpedia: 117,265 produk perangkat lunak	1. Incremental Diffusive Clustering (IDC) Association Rule Mining k-Nearest Neighbors (kNN)	IDC mampu melakukan clustering fitur dengan presisi tinggi pada dataset besar. Kombinasi antara Association Rule Mining dan kNN memberikan akurasi tinggi dalam merekomendasikan fitur
5	Recommending New Features from Mobile App Descriptions (Jiang et al., 2019)	Mengembangkan pendekatan otomatis bernama SAFER (Similar App-based FEature Recommender) untuk mengekstraksi dan merekomendasikan fitur baru berdasarkan deskripsi aplikasi.Mengatasi tantangan dalam identifikasi fitur dan identifikasi aplikasi serupa	8.359 aplikasi yang diambil dari Google Play	model SAFER (Similar App-based FEature Recommender) Reference App Filter, App Feature Extractor (AFE), API Extractor, App Profile Builder, Topic Model, Similar App Identifier, Feature Recommender	SAFER efektif dalam merekomendasikan fitur baru, SAFER mengungguli metode baseline (KNN+ dan CLAP)

SEKOLAH PASCASARJANA

Tabel 2.8 SLR Pengembangan Fitur

No	Judul	Tujuan Penelitian	Dataset	Metode Penelitian	Temuan Utama
6	A Conceptual Basis for Feature Engineerin (Reid Turner et al., 1999)	Menyatukan perspektif pengguna dan pengembang dalam pengembangan perangkat lunak melalui konsep Feature Engineering.Mengatasi kesenjangan antara domain masalah dan domain solusi dalam pengembangan perangkat lunak. Mengusulkan Feature Engineering sebagai pendekatan untuk menjadikan fitur sebagai objek utama dalam proses pengembangan perangkat lunak	Perangkat lunak switching telepon dari Italtel perangkat lunak berskala besar dengan jutaan baris kode dan ribuan file.	Feature-Oriented Development yang mencakup beberapa tahapan utama: Definisi Konsep Fitur; Model Konseptual Fitur; plikasi dalam Siklus Hidup Perangkat Lunak	Feature-Oriented Development Meningkatkan Keterlacakan Feature Isolation Meningkatkan Fleksibilitas dan Modularitas. Performa pada Studi Kasus Italtel. Identifikasi Masalah Interaksi Fitur. Evaluasi Sistem Manajemen Konfigurasi
7	Recommending Software Features for Mobile Applications Based on User Interface Comparison (Lo, 2019)	Merekomendasikan fitur perangkat lunak untuk aplikasi mobile dengan menganalisis antarmuka pengguna (UI) dari aplikasi yang sudah ada.	informasi UI dari aplikasi yang dikumpulkan dari dua sumber: 1Mobile dan F-Droid Platform untuk distribusi aplikasi open-source Android.	Perhitungan Kemiripan UI (UI Similarity Calculation). Algoritma genetika (GA) digunakan untuk mencocokkan komponen UI dan menemukan pasangan yang memiliki kesamaan optimal. Identifikasi Fitur (Feature Identification) Rekomendasi Fitur (Feature Recommendation)	Kemiripan UI Sangat Akurat: Nilai MAP (Mean Average Precision) untuk pengujian kemiripan UI adalah 0.418 (mendekati nilai optimal 0.457). Algoritma Genetika (GA) Efektif untuk Meningkatkan Efisiensi
8	Model Analytic in Fintech User Comment Features Using LDA-CNN on Imbalanced Data	Mengembangkan model analitik gabungan algoritma LDA-ROS-NCL-CNN untuk menganalisis komentar pengguna layanan aplikasi Fintech dan mengklasifikasikan topik ulasan, sehingga menghasilkan fitur yang relevan.	Google Play Komentar Pengguna Fintech di Indonesia	Metode **Probabilistik** dan Klasifikasi **Convolution** dengan penambahan imbalance data	Fitur utama dalam fintech Fitur baru Keamanan dan Modal Usaha

SEKOLAH PASCASARJANA

2.1.2 Penelitian Sebelumnya Terkait Topik Model LDA

Pemodelan topik merupakan penggalian teks untuk memproses, mengatur, mengelola, dan mengekstrak pengetahuan berdasarkan pemodelan probabilistik yang digunakan untuk menemukan struktur tersembunyi dalam arsip besar dokumen berdasarkan pola penggunaan kata yang serupa di setiap dokumen. Ini digunakan untuk menentukan topik yang mendasari dalam dokumen teks. Sebuah topik mewakili konsep yang lebih luas yang dimiliki oleh korpus dokumen dan topik ini berkembang seiring waktu (Lamba & Madhusudhan, 2019).

Algoritma topik model digunakan untuk menemukan tema utama dalam dokumen dalam koleksi dokumen yang besar dan tidak terstruktur. Topik model dapat mengatur koleksi sesuai dengan tema yang ditemukan. Algoritma ini dapat diterapkan pada koleksi dokumen yang sangat besar, dapat disesuaikan dengan berbagai jenis data (Blei et al., 2010), membantu mengidentifikasi topik dalam konteks (Lamba & Madhusudhan, 2019), dan digunakan dalam mengekstraksi dokumen produk dalam analisis sentimen berbasis aspek (Ozyurt & Akcayol, 2020). Setiap dokumen adalah campuran topik, dokumen dapat berisi kata-kata dari beberapa topik dengan proporsi tertentu. Setiap topik adalah campuran kata-kata, per kata-kata dapat dibagi di antara topik, dan kata dapat muncul di beberapa topik.

Penelitian topik model menggunakan algoritma LDA telah dilakukan di berbagai disiplin ilmu terlihat pada tabel 2.9, penelitian LDA fokus pada penemuan topik pada beberapa penelitian berikut: Penelitian (Lai & Kontokosta, 2019a) pada topik model dan analisis deret waktu, menggunakan data set property dan informasi pajak yang mampu melakukan generalisasi data teks tidak terstuktur; pendekatan LDA digunakan dalam mengekstraksi produk dalam analisis sentimen berbasis aspek (Ozyurt & Akcayol, 2020), mengidentifikasi topik utama sebagai pertimbangan perusahaan dalam kampanye pemasaran (Reyes-Menendez et al., 2020). Penelitian algoritma LDA yang dimanfaatkan memisahkan pendapat pada data komentar hotel, restoran dan perangkat elektronik multi bahasa (García-Pablos et al., 2018). Menganalisa tren topik secara otomatis pada komentar pengguna aplikasi hotel dari waktu ke waktu, menghitung, dan menganalisis net promoter

score (NPS) dari skor rating pelanggan dan menyajikannya dalam bentuk dashboard (Nguyen & Ho, 2023). Melakukan evaluasi kepercayaan secara efektif mengidentifikasi kata yang bermakna dan merepresentasikan di setiap topik pada data stack exchange (X. Chen et al., 2020b), penelitian tersebut memanfaatkan topik untuk melakukan identifikasinya. Dataset dari platform tanya-jawab (Stack Exchange) dapat digunakan untuk evaluasi kepercayaan berbasis topik, dataset dari ulasan produk dan layanan dapat digunakan untuk ekstraksi aspek dan analisis sentimen, dataset dari platform media sosial (Twitter) dapat digunakan untuk analisis opini dan strategi pemasaran, dataset dari artikel Wikipedia dapat digunakan untuk analisis perubahan data yang dinamis. Hal ini menunjukkan bahwa pola dataset yang bervariasi LDA mampu bekerja secara efektif dalam menemukan topik. Dalam penelitian ini diusulkan Algoritma LDA digunakan untuk menemukan topik pada dataset komentar pengguna Fintech untuk menemukan topik terlihat pada tabel 2.9 nomor 9.

SEKOLAH PASCASARJANA

Tabel 2.9 SLR Penelitian LDA

No	Judul	Tujuan Penelitian	Dataset yang Digunakan	Metode Penelitian	Temuan Utama
1	A topic-sensitive trust evaluation approach for users in online communities (X. Chen et al., 2020a)	Tujuan utama adalah membangun model yang mampu mengenali perbedaan konteks dan mengklasifikasikan kepercayaan berdasarkan topik tertentu.Mengembangkan Kerangka Kerja Evaluasi Kepercayaan yang Berbasis Topik	Dataset: Stack Exchange	Model User-Topic Berbasis Labeled Latent Dirichlet Allocation (LLDA) Modifikasi pada Algoritma Evaluasi Kepercayaan	Peningkatan Akurasi Evaluasi Kepercayaan Metode TS-TrustRank dan TS-Appleseed menunjukkan peningkatan akurasi dalam memprediksi kepercayaan berdasarkan topik dibandingkan dengan metode yang tidak mempertimbangkan topik.
2	2VLDA: Almost unsupervised system for Aspect Based Sentiment Analysis (García-Pablos et al., 2018)	Mengintegrasikan Pemodelan Topik dengan Embedding Kata untuk Ekstraksi Aspek dan Sentimen	SemEval 2016 Task 5 Dataset Food (Makanan) Service (Pelayanan) Ambiance (Suasana)	LDA (Latent Dirichlet Allocation) Penggunaan Word2Vec untuk Pemodelan Semantik Distribusi aspek dihitung dari distribusi topik yang diperoleh dari LDA. metode Gibbs Sampling	Peningkatan Kinerja dalam Klasifikasi Aspek dan Sentimen, Performa Lebih Baik dalam Klasifikasi Sentimen
3	Topic modeling to discover the thematic structure and spatial-temporal patterns of building renovation and adaptive reuse in cities (Lai & Kontokosta, 2019b)	Menganalisis Pola Perubahan Urban melalui Renovasi Bangunan dan Adaptasi Ulang (Adaptive Reuse) Mengembangkan Metode yang Dapat Digeneralisasi ke Berbagai Kota dan Konteks Urban	Data Izin Bangunan dari 7 Kota Besar di Amerika Serikat	Pemodelan Topik dengan LDA (Latent Dirichlet Allocation) Pemodelan Spasial dan Temporal	LDA berhasil mengidentifikasi tiga topik utama: Renovation, Addition, Change of Use
4	A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: SS-LDA"(Ozyurt & Akcayol, 2020)	Mengusulkan Model SS-LDA (Sentence Segment LDA) untuk Ekstraksi Aspek	e-commerce Hepsiburada.com (situs populer di Turki).	Pengelompokan Aspek dengan SS-LDA; Algoritma Gibbs Sampling; Deteksi Sentimen	Kinerja Model dalam Ekstraksi Aspek Generalitas dan Fleksibilitas

SEKOLAH PASCASARJANA

Tabel 2.9 SLR Penelitian LDA lanjutan

No	Judul	Tujuan Penelitian	Dataset yang Digunakan	Metode Penelitian	Temuan Utama
5	Analysing online customer experience in hotel sector using dynamic topic modelling and net promoter score (Nguyen & Ho, 2023)	Memahami Pengalaman Pelanggan di Industri Perhotelan; Menggabungkan Model Dinamis untuk Menganalisis Perubahan Pola Topik Seiring Waktu	Data ulasan pelanggan diambil dari situs Agoda (platform ulasan hotel).	Pemodelan Topik dengan LDA menganalisis perubahan distribusi topik seiring waktu	Identifikasi Topik Utama dalam Pengalaman Pelanggan Hasil model LDA mengidentifikasi 15 topik utama Perubahan Pola Topik Seiring Waktu
6	A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: SS-LDA"(Ozyurt & Akcayol, 2020)	Mengusulkan Model SS-LDA (Sentence Segment LDA) untuk Ekstraksi Aspek	e-commerce Hepsiburada.com (situs populer di Turki).	Pengelompokan Aspek SS-LDA; Algoritma Gibbs Sampling; Deteksi Sentimen	Kinerja Model dalam Ekstraksi Aspek Generalitas dan Fleksibilitas
7	Marketing challenges in the #MeToo era: gaining business insights using an exploratory sentiment analysis (Reyes-Menendez et al., 2020)	Menganalisis Tantangan Bisnis dan Pemasaran dalam Konteks Gerakan #MeToo Memberikan Wawasan untuk Peningkatan Strategi Pemasaran dan Komunikasi	Data diambil dari platform media sosial Twitter. 42.351 tweet	Pemodelan Topik dengan LDA (Latent Dirichlet Allocation) Analisis Sentimen dengan SVM (Support Vector Machine)	Identifikasi 7 Topik Utama Terkait #MeToo Sentimen Terkait Setiap Topik Krippendorff's Alpha Value (KAV) untuk klasifikasi sentimen: Positif 0.697 Netral 0.754 Negatif 0.734
8	Summarization of changes in dynamic text collections using Latent Dirichlet Allocation model (Kar et al., 2015)	Mengembangkan Sistem untuk Merangkum Perubahan pada Koleksi Dokumen Teks Dinamis	Dataset terdiri dari 54 artikel Wikipedia yang dipilih secara manual	Model Pemeringkatan Kalimat Penerapan LDA (Latent Dirichlet Allocation)	Model LDA menghasilkan 5 topik laten
9	Model Analytic in Fintech User Comment Features Using LDA-CNN on Imbalanced Data	Mengembangkan model analitik gabungan algoritma LDA-ROS-NCL-CNN untuk menganalisis komentar pengguna layanan aplikasi Fintech dan mengklasifikasikan topik ulasan, sehingga menghasilkan fitur yang relevan.	Google Play Komentar Pengguna Fintech di Indonesia	Metode Probabilistik LDA Coherensi, Jaccard Similarity dan Klasifikasi CNN dengan penambahan imbalance data	Menghasilkan 10 fitur

SEKOLAH PASCASARJANA

Manfaat klasifikasi teks adalah untuk otomaisi kategori teks ke dalam kelas yang telah ditentukan sebelumnya, merespons pertanyaan pengguna yang bertujuan memahami teks (Kamruzzaman et al., 2010), meningkatkan ketepatan sistem pengambilan informasi dengan memastikan hanya informasi relevan yang diambil sesuai dengan kebutuhan pengguna (Nithya et al., 2012), pemfilteran spam, identifikasi bahasa, analisis sentiment (Vijayan et al., 2017). Sistem klasifikasi teks modern dirancang agar dapat beradaptasi dengan berbagai jenis data dan cukup akurat untuk memenuhi harapan pengguna, sehingga memberikan manfaat bagi berbagai domain termasuk manajemen konten web dan penanganan data perusahaan (X. Zhou et al., 2020).

LDA berhasil meningkatkan klasifikasi dokumen multi-label dan menunjukkan peningkatan performa yang signifikan (X. Li et al., 2015). Penelitian (Ramage et al., 2009) fokus pada pengembangan LDA untuk klasifikasi multi-label dengan hasil yang lebih akurat melalui Gibbs Sampling. Penelitian Mehrotra et al., 2013 fokus pada peningkatan koherensi topik pada data Twitter dengan strategi pooling dan pelabelan otomatis. Penelitian (Wahid et al., 2022) fokus pada klasifikasi berbasis konteks untuk data krisis menggunakan kombinasi LDA, TF-IDF. Penelitian ini proses pelabelan diusulkan berdasarkan bobot setiap kata kunci hasil proses LDA yang ada dalam setiap topik. Dengan melakukan ekstraksi distribusi topik dalam dokumen, mengurutkan topik berdasarkan kontribusi, menentukan topik dominan, mengambil kata kunci dalam topik dominan, menghitung persentase kontribusi, melabeli dokumen berdasarkan topik dominan.

SEKOLAH PASCASARJANA

Tabel 2.10 SLR Pelabelan berdasarkan Topik LDA

No	Judul	Tujuan Penelitian	Dataset yang Digunakan	Metode Penelitian	Temuan Utama
1	Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-Labeled Corpora (Ramage et al., 2009)	Mengembangkan model LDA yang dapat menangani multi-label	Diambil dari situs bookmark sosial del.icio.us 4000 dokumen dengan setidaknya satu dari 20 tag	Model Labeled LDA Setiap topik dalam L-LDA langsung dikaitkan dengan label yang telah diberikan pada dokumen. Gibbs Sampling	Labeled LDA menghasilkan cuplikan yang lebih relevan dibandingkan SVM
2	Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling (Mehrotra et al., 2013)	Meningkatkan Koherensi Topik pada LDA di Data Twitter Mengembangkan Skema Pelabelan Hashtag Otomatis	Generic Dataset (359.478 tweet) – Dataset umum dengan kata kunci seperti musik, bisnis, dan film	TF-IDF Model LDA dilatih pada data hasil pooling	Hashtag-based Pooling memberikan hasil terbaik untuk semua matrik evaluasi: PMI meningkat hingga +164% dibanding metode dasar. Automatic Hashtag Labeling meningkatkan hasil lebih lanjut:
3	Document Classification by Topic Labeling (Hingmire et al., 2013)	Mengembangkan algoritma klasifikasi dokumen berbasis LDA; Mengurangi ketergantungan pada dataset berlabel; Mengatasi etidakakuratan dalam proses pelabelan otomatis	20Newsgroup Berisi pesan dari 20 grup diskusi	LDA (Latent Dirichlet Allocation) untuk Klasifikasi Topik Ekstensi Algoritma dengan EM dan Naive Bayes (ClassifyLDA-EM)	Mengembangkan metode klasifikasi berbasis LDA dengan pelabelan otomatis. Meningkatkan akurasi klasifikasi menggunakan EM dan Naive Bayes.
4	Topic2Labels: A Framework to Annotate and Classify the Social Media Data Through LDA Topics and Deep Learning Models (Wahid et al., 2022)	Mengembangkan Pendekatan Label Otomatis. Menerapkan LDA untuk menghasilkan topik dari data dan menentukan label dominan untuk setiap dokumen.	Dataset berisi 10 juta tweet terkait pandemi COVID-19 dari 1 Maret hingga 30 April 2020.	Topic Modeling menggunakan LDA Topik Dominan (T-TF-IDF)	Mengembangkan metode pelabelan otomatis berbasis LDA dan TF-IDF; LSTM menghasilkan hasil yang lebih baik dibanding ANN karena mampu menangkap pola konteks dengan lebih baik

SEKOLAH PASCASARJANA

2.1.3 Penelitian Sebelumnya Terkait Imbalance Data

Ketidakseimbangan data (imbalanced data) terjadi ketika jumlah sampel antar kelas dalam dataset tidak merata (Rasheed et al., 2022) (Rasheed et al., 2022), Strategi oversampling dan undersampling, digunakan untuk menangani ketidakseimbangan kelas dalam pembelajaran mesin. Dataset dianggap imbalanced jika kelas minoritas memiliki ≤ 25% dari total instance dalam dataset (Barella et al., 2021). Dalam penanganan data teks yang tidak seimbang, penggunaan metode RandomOverSampler (ROS) dan Neighbourhood Cleaning Rule (NCL) menjadi strategi dalam penyelesaian masalah ini. Hybrid metode undersampling dan oversampling diusulkan untuk mengatasi ketidakseimbangan data. Beberapa penelitian yang menggunakan ROS dan NCL secara terpisah antara lain metode NCL sebagai metode undersampling mampu meningkatkan nilai akurasi (Saadah & Wulandari, 2015; Agustianto & Destarianto, 2019). Metode ROS memiliki kinerja yang lebih baik sebagai opsi untuk mengatasi data tidak seimbang (Kamalov et al., 2023). ROS secara konsisten mengungguli under-sampling (RUS) (Johnson & Khoshgoftaar, 2020a). NCL mampu meningkatkan pemodelan kelas kecil yang sulit untuk membangun pengklasifikasi untuk mengidentifikasi kelas-kelas kecil (Laurikkala, 2001).

Penelitian terdahulu melakukan kombinasi teknik dalam menangani data tidak seimbang terlihat pada tabel 2.11, penelitian. Gabungan teknik Borderline-SMOTE menunjukkan hasil yang efektif dalam meningkatkan performa klasifikasi pada 203 dataset diambil dari platform OpenML (Rasheed et al., 2022), ROS-RUS memberikan hasil terbaik dalam ketidakseimbangan data pada big data CMS Medicare (Johnson & Khoshgoftaar, 2019), teknik hybrid (SMOTEENN dan SVMSMOTE) memberikan performa tertinggi pada dataset tidak seimbang 2 dataset sintetis yang berasal dari CERT (Jaiswal et al., 2024), Teknik hybrid ROS-RUS meningkatkan kecepatan pelatihan dan performa Data pada data klaim layanan kesehatan di Amerika Serikat dari tahun 2012–2016 (Johnson & Khoshgoftaar, 2020a), kombinasi NCL-SMOTE secara efektif meningkatkan recall pada dataset tidak seimbang pada 5 dataset medis (Junsomboon & Phienthrakul, 2017).

Berdasarkan penelitian terdahulu tersebut, penelitian ini diusulkan untuk meningkatkan performa prediksi dalam klasifikasi data tidak seimbang dengan menggabungkan dua teknik utama ROS dan NCL. Kombinasi teknik dalam menangani dataset tidak seimbang sangat penting karena setiap teknik memiliki kelebihan masing-masing. Dengan menggabungkan dua teknik dapat memanfaatkan kelebihan dari masing-masing metode dan mengurangi kelemahan yang muncul jika metode tersebut digunakan secara independen.

2.1.4 Penelitian Sebelumnya Terkait Klasifikasi CNN

Penelitian terdahulu algoritma CNN pada tabel 2.12, dijelaskan bahwa CNN yang digabungkan dengan algoritma lain dapat meningkatkan akurasi dan memperkaya informasi. Penelitian semantic extension: Pendekatan dengan memperluas makna semantik (Set-CNN) mampu meningkatkan akurasi karena memperkaya informasi yang digunakan dalam klasifikasi hybrid CNN-LSTM, kombinasi CNN untuk fitur lokal dan LSTM untuk ketergantungan konteks mampu meningkatkan akurasi secara signifikan. (Y. Zhou et al., 2022). Penggunaan Attention CNN-Attention memperbaiki pemahaman konteks dalam klasifikasi teks berbasis domain tertentu. (Banerjee et al., 2019). Penggabungan Binary Relevance-CNN menunjukkan hasil yang baik dalam klasifikasi multi-label, terutama dengan model deep learning berbasis CNN. (Yang & Emmert-Streib, 2024). Penggabungan CNN -LSTM meningkatkan kemampuan dalam menangkap fitur pada berbagai level kata, frasa, dan kalimat (X. Li & Ning, 2020),

Berdasarkan kemampuan algoritma CNN dalam meningkatkan kemampuan dalam menangkap fitur dan kemampuan CNN dalam penggabungan algoritma, maka penelitian ini mengusulkan algoritma CNN dengan menggabungkan algoritma LDA, ROS, NCL untuk klasifikasi fitur untuk membangun model yang baik dalam menangani komentar pengguna Fintech.

Tabel 2.11 SLR Penanganan Data Tidak Seimbang

No	Judul	Tujuan Penelitian	Dataset yang Digunakan	Metode Penelitian	Temuan Utama
1	Assessing the Data Complexity of Imbalanced Datasets (Rasheed et al., 2022)	Menyelidiki efektivitas ukuran kompleksitas data pada dataset tidak seimbang dan bagaimana pengaruhnya setelah menerapkan perlakuan ketidakseimbangan data	203 dataset diambil dari platform OpenML untuk analisis eksperimen.	Menganalisis perubahan kompleksitas dataset sebelum dan sesudah menerapkan teknik DIT RO, RUS, SMOTE, Borderline-SMOTE ADASYN, CBO, (OSS)	Teknik DIT mampu mengurangi kompleksitas dataset dan meningkatkan performa prediksi secara signifikan, terutama pada dataset dengan ketidakseimbangan tinggi. Teknik SMOTE dan Borderline-SMOTE menunjukkan hasil yang paling efektif dalam meningkatkan performa klasifikasi pada dataset tidak seimbang.
2	"Deep Learning and Data Sampling with Imbalanced Big Data (Johnson & Khoshgoftaar, 2019)	Menilai efektivitas metode deep learning dalam menangani masalah ketidakseimbangan kelas pada big data untuk deteksi kecurangan Medicare. Menganalisis pengaruh berbagai metode sampling terhadap pelatihan dan performa model deep learning, termasuk: ROS, RUS Hybrid ROS-RUS	Dataset yang digunakan adalah CMS Medicare Part B (2012–2016) yang disediakan oleh Centers for Medicare and Medicaid Services (CMS).	Arsitektur Deep Neural Network (DNN) Teknik Sampling RUS, ROS, Hybrid ROS-RUS	Menunjukkan bahwa ROS-RUS memberikan hasil terbaik dalam ketidakseimbangan data pada big data. Teknik hybrid (ROS-RUS) memungkinkan pelatihan cepat dan performa tinggi.
3	Handling imbalance dataset issue in insider threat detection using *machine learning* methods (Jaiswal et al., 2024)	Mengatasi masalah ketidakseimbangan pada dataset dalam deteksi ancaman orang dalam. Mengembangkan pendekatan baru yang disebut Two-Step Insider Threat Detection (TSITD	2 dataset sintetis yang berasal dari CERT (Computer Emergency Response Team) yang dibuat oleh Software Engineering Institute (SEI) di Carnegie Mellon University	Digunakan 8 algoritma klasifikasi untuk deteksi ancaman orang dalam: ANN, RVFL, ELM RF, DT, XGB, kNN, SVM	Mengusulkan kerangka kerja TSITD untuk mengatasi ketidakseimbangan data dalam deteksi insider threat. Membuktikan bahwa teknik hybrid (SMOTEENN dan SVMSMOTE) memberikan performa tertinggi pada dataset tidak seimbang.

SEKOLAH PASCASARJANA

Tabel 2.11 SLR Penanganan Data Tidak Seimbang lanjutan

No	Judul	Tujuan Penelitian	Dataset yang Digunakan	Metode Penelitian	Temuan Utama
4	The Effects of Data Sampling with Deep Learning and Highly Imbalanced Big Data (Johnson & Khoshgoftaar, 2020a)	Mengevaluasi efektivitas metode data sampling dalam menangani ketidakseimbangan kelas pada big data menggunakan DNNs. Menganalisis pengaruh ROS, RUS, dan kombinasi keduanya ROS-RUS terhadap performa model pada tiga dataset besar dengan ketidakseimbangan tinggi.	Medicare Part B – Data klaim layanan kesehatan di Amerika Serikat dari tahun 2012–2016.	Arsitektur Deep Neural Network (DNN) Random Under-Sampling (RUS) Random Over-Sampling (ROS): Hybrid ROS-RUS	ROS-RUS memberikan kombinasi terbaik antara performa dan efisiensi. Teknik hybrid ROS-RUS meningkatkan kecepatan pelatihan dan performa.
5	Combining Over-Sampling and Under-Sampling Techniques for Imbalance Dataset (Junsomboon & Phienthrakul, 2017)	Mengatasi masalah ketidakseimbangan data pada data medis. Meningkatkan performa prediksi dalam klasifikasi data medis dengan cara menggabungkan teknik NCL-SMOTE.	5 dataset medis yang dibagi menjadi training set (67%) dan testing set (33%)	Menggunakan kombinasi metode NCL dan SMOTE untuk menyeimbangkan	Kombinasi NCL + SMOTE secara efektif meningkatkan recall pada dataset medis yang tidak seimbang. Metode ini efektif karena NCL menghilangkan outlier dan SMOTE menambahkan sampel sintetis untuk menyeimbangkan data.

SEMARANG
SEKOLAH PASCASARJANA

Tabel 2.12 SLR Klasifikasi CNN

No	Judul	Tujuan Penelitian	Dataset yang Digunakan	Metode Penelitian	Temuan Utama
1	Set-CNN: A text convolutional neural network based on semantic extension for short text classification (Y. Zhou et al., 2022)	enelitian ini bertujuan untuk mengembangkan (Semantic Extension Text-CNN, sebuah algoritma klasifikasi teks pendek berbasis perluasan semantic Meningkatkan klasifikasi teks pendek; Menangkap informasi semantic; Mengurangi kompleksitas komputasi	5000 kalimat subjektif dan 5000 kalimat objektif untuk tugas klasifikasi subjektif atau objektif. 5952 pertanyaan berlabel, 9613 kalimat yang diambil dari ulasan film. Mandarin yang terdiri dari 63.831 sampel dalam enam kelas teks pendek (seperti pendidikan, kesehatan, dll.).	Perluasan Semantik (Semantic Extension); Lapisan Konvolusi Multi-Channel; Multiple-channel Text-CNN; K-Max Pooling	Set-CNN berhasil meningkatkan performa klasifikasi teks pendek dengan: Perluasan semantik yang efektif untuk mengatasi kelangkaan fitur pada teks pendek. Kombinasi konvolusi multi-channel yang menangkap informasi semantik lokal dan global. Efisiensi komputasi yang tinggi, menjadikannya alternatif yang ringan namun tetap kompetitif dengan model besar seperti BERT.
2	Comparative effectiveness of CNN and RNN architectures for radiology text report lassification (Banerjee et al., 2019)	Membandingkan efektivitas model deep learning berbasis CNN dan RNN dalam klasifikasi laporan radiologi berbasis teks bebas. CNN Word-Glove – Model CNN Hierarchical Recurrent Neural Network (DPA-HNN	Stanford University Medical Center 117.816 laporan radiologi tentang pasien dewasa.	CNN Word – Glove Architectur Domain Phrase Attention-based DPA-HNN	Model DPA-HNN menunjukkan kinerja terbaik pada sebagian besar pengujian. Model yang dilatih pada data Stanford menunjukkan kemampuan generalisasi tinggi pada dataset eksternal
3	"Optimal performance of Binary Relevance CNN in targeted multi-label text classification (Yang & Emmert-Streib, 2024)	Menyelidiki efektivitas Binary Relevance (BR) yang berbasis deep learning pada tugas klasifikasi teks multi-label (MLTC). BR-CNN – Convolutional Neural Network untuk klasifikasi berbasis BR.	Arxiv Academic Paper Dataset (AAPD) euters-21578 Reuters Corpus Volume I (RCV1-v2)	Model Binary Relevance (BR) yang Diuji BR-CNN; BR-LSTM; BR-GRU BR-SVM; BR-decoder – Model RNN	BR-CNN menunjukkan kinerja terbaik untuk AAPD dan MIMIC-III BR-CNN mampu mengungguli metode multi-label yang menggunakan informasi ketergantungan label pada dataset tertentu.

SEKOLAH PASCASARJANA

Tabel 2.12 SLR Klasifikasi CNN Lanjutan

No	Judul	Tujuan Penelitian	Dataset yang Digunakan	Metode Penelitian	Temuan Utama
4	Chinese text classification based on hybrid model of CNN and LSTM (X. Li & Ning, 2020)	Mengembangkan model Convolutional Neural Network (CNN) untuk tugas klasifikasi teks.	Tidak tertulis	Arsitektur CNN untuk Klasifikasi Teks Pengaturan Hyperparameter	CNN menunjukkan peningkatan akurasi yang signifikan dibandingkan dengan metode baseline lainnya. Kombinasi beberapa ukuran kernel memungkinkan CNN menangkap pola kata, frasa, dan kalimat yang berbeda dalam teks.
5	Text Classification Based on Hybrid CNN-LSTM Hybrid Model(She & Zhang, 2018)	Mengusulkan model hibrid CNN-LSTM untuk meningkatkan akurasi klasifikasi teks dengan: CNN untuk mengekstraksi fitur lokal dari teks. LSTM untuk menangkap ketergantungan konteks. Kombinasi fitur dari CNN sebagai input ke LSTM.	Sogou Chinese News Corpus	Word2Vec Layer CNN LSTM	Model hibrid CNN-LSTM menunjukkan kinerja lebih baik daripada metode konvensional seperti KNN, SVM, CNN, dan LSTM. Kombinasi CNN dan LSTM mampu meningkatkan akurasi karena menangkap fitur lokal dan konteks sekaligus.
6	Analysis and Research on Automatic Text Classification Algorithm Based on *Machine learning* (Xie & Luo, 2024)	Mengatasi permasalahan dalam pengelolaan data teks yang besar di era big data. Mengembangkan metode klasifikasi teks otomatis untuk menggantikan metode klasifikasi manual yang membutuhkan banyak waktu dan tenaga.	THUCNews dari Tsinghua University, dengan detail sebagai berikut: Periode Data: Data berasal dari RSS feed berita Sina News antara tahun 2005 hingga 2011.	Plain Bayesian Algorithm; CNN; LSTM	CNN memiliki akurasi tertinggi dan kemampuan generalisasi yang baik. Plain Bayesian memiliki performa yang kurang baik pada data kompleks. LSTM mampu menangkap konteks tetapi kalah efisien dibandingkan CNN.

SEKOLAH PASCASARJANA

2.2 Keaslian Penelitian

Penelitian Fintech memiliki area riset yang luas dengan dataset yang bervariasi. Namun dataset teks komentar pengguna fintech memiliki potensi untuk dikembangkan dalam penelitian Fintech. Dataset teks komentar pengguna Fintech untuk pengembangan model berbasis *Machine learning* merupakan area riset yang masih belum dieksplorasi. Penelitian ini membawa perspektif baru dalam memanfaatkan komentar pengguna Fintech sebagai alternatif data yang inovatif.

Saat ini belum ada penggunaan algoritma LDA untuk mengidentifikasi fitur Fintech berdasarkan komentar pengguna yang dapat dimanfaatkan untuk pengembangan layanan aplikasi. Maka diusulkan topik yang dihasilkan digunakan untuk proses pelabelan dokumen. Dan topik yang dihasilkan akan dilakukan pencocokan kata menggunakan OpenAI berdasarkan fitur untuk memakai topik ke dalam fitur Fintech.

Pada penelitian sebelumnya hasil pelabelan topik menggunakan algoritma LDA belum ada yang membahas tentang data tidak seimbang, sementara kumpulan data yang tidak seimbang merupakan masalah yang sering ditemukan dan ketidakseimbangan data merupakan masalah utama yang dihadapi saat menerapkan konsep pembelajaran mesin (Yu & Zhou, 2021) yang tidak diawasi maupun pembelajaran yang diawasi (Budania et al., 2020). Penggabungan algoritma LDA - CNN dengan penambahkan ROS-NCL untuk mengisi kesenjangan pada proses pra pengolahan dan klasifikasi untuk meningkatkan kinerja data berlabel yang masih kurang optimal.

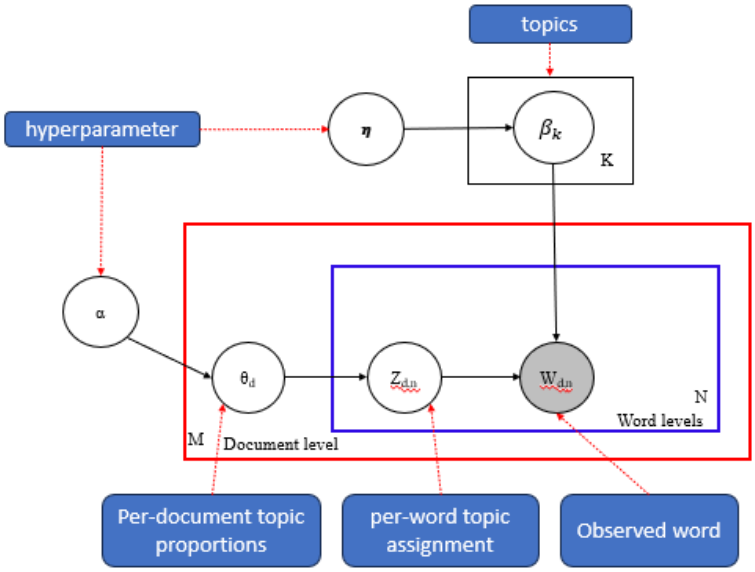
Algoritma *machine learning* dan *deep learning* telah banyak digunakan dalam pengolahan teks, namun penggabungan algoritma LDA-ROS-NCL-CNN pada Komentar pengguna Fintech belum ada, maka diusulkan penggabungan algoritma tersebut untuk membangun model analitik fintech

SEKOLAH PASCASARJANA

2.3 Landasan Teori

2.3.1 Topik Model LDA

LDA merupakan model generatif probabilistik yang dapat digunakan untuk memperkirakan observasi multinomial dengan pendekatan pembelajaran tanpa pengawasan, untuk memperkirakan menemukan campuran kata-kata yang terkait dengan setiap topik, dan menentukan campuran topik yang mendeskripsikan setiap dokumen. Di bawah ini gambaran 2.2 proses topik model menggunakan LDA.



Gambar 2. 1 Diagram Model LDA

Dengan:

M menunjukkan jumlah dokumen

N adalah jumlah kata dalam dokumen tertentu (dokumen i memiliki N_i kata)

K - Jumlah topik dalam model,menggambarkan bahwa ada beberapa topik yang mungkin, dan setiap topik memiliki distribusi kata terkait.

α: Parameter hiperprior untuk distribusi Dirichlet dari distribusi topik pada dokumen, θ.

θ_d:Distribusi topik pada dokumen d, diambil dari distribusi Dirichlet dengan parameter α.

$Z_{d,n}$: Topik untuk kata ke-n dalam dokumen d, dipilih berdasarkan distribusi topik θ_d

SEMARANG
SEKOLAH PASCASARJANA

$W_{d,n}$: Kata ke-n dalam dokumen d, dipilih berdasarkan topik $Z_{d,n}$ dan distribusi kata ϕk yang berkaitan dengan topik tersebut.
β_k:Parameter hiperprior untuk distribusi Dirichlet dari distribusi kata pada topik, ϕ.
η: parameter Dirichlet untuk distribusi kata dalam topik.

LDA melakukan proses pengolahan teks dengan rumus 2.1 dan 2.2 (Nguyen & Ho, 2023). Rumus 1 melakukan distribusi kata per topik dan distribusi topik per dokumen, penentuan topik untuk setiap kata dalam setiap dokumen, pengamatan kata berdasarkan distribusi kata dalam topik yang relevan.

$$p(\beta_{1:K}, \theta_{1:D}, Z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(Z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, Z_{d,n})\right) \quad (2.1)$$

$\beta_{1:K}$: Parameter yang mewakili distribusi kata-kata untuk setiap topik. βi adalah vektor probabilitas kata-kata untuk topik ke-i.
$\theta_{1:D}$: Parameter yang mewakili distribusi topik untuk setiap dokumen. θd adalah vektor probabilitas topik untuk dokumen ke-d.
$z_{1:D}$: Variabel laten yang menentukan topik dari setiap kata dalam setiap dokumen.
$w_{1:D}$: Kata-kata yang diamati dalam dokumen.

$\prod_{i=1}^{K} p(\beta_i)$: Probabilitas dari distribusi topik. Ini menunjukkan distribusi kata-kata untuk setiap topik dalam korpus, ini adalah distribusi Dirichlet.
$\prod_{d=1}^{D} p(\theta_d)$: Probabilitas dari distribusi topik per dokumen. Ini menunjukkan distribusi topik yang ada dalam setiap dokumen, menggunakan distribusi Dirichlet.
$\prod_{n=1}^{N} p(Z_{d,n}|\theta_d)$: Probabilitas dari penugasan topik untuk setiap kata dalam dokumen, mengingat distribusi topik dari dokumen tersebut.
$\prod_{n=1}^{N} p(w_{d,n} \mid \beta_{1:K}, Z_{d,n})$: Probabilitas dari kata yang diamati dalam dokumen, mengingat penugasan topik dan distribusi kata per topik.

Proses pada rumus 2.2 digunakan untuk menghitung distribusi posterior dari parameter model dan variabel laten berdasarkan data kata yang diamati.

SEKOLAH PASCASARJANA

$$p(\beta_{1:K}, \theta_{1:D}, Z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K},\theta_{1:D},Z_{1:D},W_{1:D})}{p(W_{1:D})} \quad (2.2)$$

$\beta_{1:K}$: Parameter yang mewakili distribusi kata-kata untuk setiap topik.
$\theta_{1:D}$: Parameter yang mewakili distribusi topik untuk setiap dokumen.
$z_{1:D}$: Variabel laten yang menentukan topik dari setiap kata dalam setiap dokumen.
$w_{1:D}$: Kata-kata yang diamati dalam dokumen.

Posterior Distribution $p(\beta_{1:K}, \theta_{1:D}, Z_{1:D}|w_{1:D})$ adalah distribusi probabilitas dari parameter model dan variabel laten berdasarkan kata-kata yang diamati. Distribusi ini memberikan pemahaman tentang topik yang paling mungkin menghasilkan setiap kata dalam dokumen setelah mempertimbangkan kata-kata yang diamati.

Joint Distribution $p(\beta_{1:K}, \theta_{1:D}, Z_{1:D}, W_{1:D}$, adalah probabilitas gabungan dari semua parameter model, variabel laten, dan kata-kata yang diamati. Fungsi ini menggambarkan probabilitas terjadinya seluruh pengaturan ini dalam korpus.
Marginal Probability $p(W_{1:D})$, adalah probabilitas dari kata-kata yang diamati tanpa mempertimbangkan penugasan topik atau distribusi topik tertentu. Ini merupakan penjumlahan atas semua kemungkinan konfigurasi parameter dan variabel laten.

Rumus 2.1 digunakan untuk menghitung probabilitas gabungan dari semua komponen model (data dan parameter) sebagai model generatif, rumus 2.2 menggunakan output dari rumus 2.1 sebagai pembilang dan menggabungkannya dengan penyebut untuk menghitung distribusi posterior. Distribusi ini untuk inferensi statistika yang memungkinkan untuk memperkirakan parameter model setelah mengamati data nyata.

SEKOLAH PASCASARJANA

Rumus 2.1 merupakan cara data bisa dihasilkan mengikuti asumsi model, sedangkan Rumus 2.2 digunakan untuk memperbarui pemahaman tentang parameter model setelah mengamati data. Rumus 2.1 adalah model dasar LDA yang menjelaskan bagaimana data (kata-kata dalam dokumen) bisa terbentuk jika diketaui beberapa parameter tertentu:

Topik dari kata-kata yaitu menggambarkan topik mana yang mungkin menghasilkan kata tertentu. Kata-kata dalam topik yaitu distribusi kata-kata untuk setiap topik. Topik dalam dokumen yaitu distribusi topik untuk setiap dokumen. Model ini adalah cara membuat sebuah dokumen dengan mengambil topik-topik tertentu dan kata-kata dari topik tersebut.

Rumus 2.2: Inferensi posterior digunakan ketika memiliki dokumen (kata-kata yang diamati) untuk mengetahui apa topik dari kata-kata tersebut, bagaimana distribusi kata-kata dalam topik, bagaimana distribusi topik dalam dokumen. Rumus ini menggunakan teorema Bayes untuk membalikkan model generatif (Rumus 2.1) agar bisa memperkirakan parameter model dari data yang ada.

Hubungan Antara Keduanya adalah rumus 2.1 memberikan cara teoretis dokumen bisa terbentuk dari topik-topik tertentu, rumus 2.2 menggunakan informasi dari dokumen yang sudah ada untuk mengetahui topik-topik dan distribusi kata dalam topik tersebut.

2.3.2 Pelabelan Topik

Topik digunakan untuk untuk melabeli dokumen diusulkan dengan menyediakan cara otomatis untuk memberi label data melalui pendekatan pemodelan topik LDA. Beberapa penelitian yang berhubungan dengan topik dan pelabelan antara lain: Model topik terawasi untuk klasifikasi multi-label dengan memasukkan frekuensi label dan relasi antar label, dua model khusus yaitu memodifikasi LDA, memberikan hasil performa yang signifikan (X. Li et al., 2015). Model LDA memanfaatkan informasi label melalui distribusi Dirichlet, efektif merumuskan hubungan antar label, yang menghasilkan peningkatan signifikan dalam klasifikasi baik label tunggal maupun multi-label (Y. Wang et al., 2020).

Saat ini OpenAI telah dimanfaatkan oleh peneliti untuk menjelaskan sebuah topik (Giray, 2023). OpenAI memberikan deskripsi yang komprehensif dan sistematis tentang metode rekayasa perintah berdasarkan model visual, yang menawarkan wawasan berharga bagi peneliti dalam melakukan eksplorasi. Penelitian memanfaatkan OpenAI telah dilakukan oleh beberapa peneliti yaitu

(Nazir et al., 2023) yang berhasil menangkap pola dan ketergantungan yang rumit dalam data biologis, yang menghasilkan imputasi yang tepat, kosakata biolog sebagai penerjemah meningkatkan akurasi dan relevansi imputasi. Large Language Models (LLM) secara konsisten menunjukkan kemampuan generalisasi saat diterapkan pada berbagai tugas bahasa (Z. Wang et al., 2023).

2.3.3 Skor Koherensi

Model topik LDA dievaluasi dengan mengamati skor koherensi untuk menemukan jumlah topik yang optimal dengan cara menghitung coherence score pada tiap model LDA yang dihasilkan dengan jumlah topik yang berbeda-beda. Skor koherensi tidak hanya memfasilitasi perbandingan model atau konfigurasi yang berbeda namun juga mengoptimalkan model untuk memastikan bahwa topik yang diekstraksi bermakna dan dapat diinterpretasikan. Penelitian menggunakan LDA menunjukkan pentingnya memilih teknik pemodelan topik dan ukuran koherensi yang tepat untuk dataset tertentu. Studi menekankan peran pembobotan istilah dan kesesuaian metode pemodelan topik yang berbeda untuk menganalisis domain non-mainstream (O'Callaghan et al., 2015). Penelitian menyelidiki dampak penggunaan abstrak versus teks lengkap terhadap koherensi topik yang dihasilkan LDA, menunjukkan bahwa koherensi dan peringkat topik LDA dipengaruhi oleh jenis data tekstual, dengan frekuensi dokumen, panjang kata, dan ukuran kosa kata menunjukkan efek beragam pada koherensi topik (Syed & Spruit, 2017). Mengeksplorasi kinerja berbagai teknik pemodelan topik pada ulasan produk e-commerce, menggunakan skor koherensi sebagai ukuran kualitas topik, mengungkapkan bahwa LDA mengungguli teknik lain pada kumpulan data mereka, maka skor koherensi sebagai ukuran penting efektivitas model topik (Chehal et al., 2021). Analisis pengalaman pelanggan online di sektor hotel menggunakan pemodelan topik dinamis (Nguyen & Ho, 2023).

Algoritma pemodelan topik LDA sering digunakan dalam analisis pelanggan (Barravecchia et al., 2022), namun tidak menyebutkan berapa topik yang ada dalam korpus tersebut. Jika terlalu sedikit topik yang dipilih, akan menyebabkan hilangnya informasi pada topik yang terlalu luas. Sebaliknya, jika terlalu banyak

SEKOLAH PASCASARJANA

topik yang diekstraksi, maka akan diperoleh topik-topik tidak berguna yang memiliki banyak kesamaan (Gan & Qi, 2021). Maka dalam model LDA diuji dengan menggunakan koherensi. Koherensi digunakan untuk mengukur kemampuan interpretasi topik, matrik koherensi topik berdasarkan asumsi bahwa jika suatu topik dapat diinterpretasikan, pasangan kata teratas dalam topik ini akan muncul lebih sering dalam dokumen korpus (B. Liu et al., 2020).

Koherensi topik C_v mengukur kesamaan semantik antar kata-kata yang sering muncul bersama dalam satu topik. Koherensi ini dihitung dengan menggabungkan poin-poin penting probabilitas bersyarat dan dukungan dokumentasi dari kata-kata tersebut. Kriteria koherensi dengan nilai 0 sampai 1, dengan tertinggi 1 dianggap sebagai angka paling koheren. Penelitian (van Dijk et al., 2023) menggunakan koherensi topik nilai tertinggi untuk mendapatkan nilai K. Dalam beberapa kasus, perhitungan koherensi didasarkan pada 10 kata teratas (Syed & Spruit, 2017). C_v didasarkan pada empat bagian: (i) segmentasi data menjadi pasangan kata, (ii) perhitungan probabilitas kata atau pasangan kata, (iii) perhitungan ukuran konfirmasi yang mengukur seberapa kuat satu set kata mendukung set kata lain, dan (iv) agregasi ukuran konfirmasi individual menjadi skor koherensi keseluruhan. Langkah-langkah penggunaan koherensi topik C_v sebagai berikut:

1. Menentukan kata-kata teratas dari setiap topik, dari topik dari model LDA, pilih N kata teratas berdasarkan probabilitas yang muncul dalam topik. Kata-kata tersebut representasi dari topik tersebut.
2. Menghitung frekuensi dokumen, untuk setiap kata w_i yang dipilih $D(w_i)$: Jumlah dokumen di korpus yang mengandung kata W_i. Untuk setiap pasangan kata w_i, w_j dengan i≠j. $D(w_i,w_j)$ adalah jumlah dokumen yang mengandung kedua kata w_i dan w_j.
3. Rumus Koherensi untuk Setiap Pasangan Kata, untuk pasangan kata (w_i,w_j), koherensi dihitung dengan rumus berikut:

$$Score(w_i, w_j) = \log\left(\frac{\mathrm{D}(w_i,w_j)+\epsilon}{\mathrm{D}(w_i)x\, D(w_j)}\right) \quad (2.3)$$

dengan ϵ adalah konstanta kecil (misalnya 10^{-12}) untuk menghindari log dari nol.

SEKOLAH PASCASARJANA

4. Agregasi skor koherensi, agar menghasilkan satu skor koherensi untuk satu topik, agregasi semua skor dari setiap pasangan kata dalam topik

$$C_v = \sum_{i<j} score(w_i, w_j) \quad (2.4)$$

Jaccard Similarity digunakan untuk mengukur tingkat kesamaan antara kata-kata di dua topik berbeda (Costa, 2021; X. Li & Zhao, 2023)

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (2.5)$$

$A \cap B$ adalah jumlah kata yang sama antara dua topik(irisan).
$A \cup B$ adalah total kata unik dari kedua topik (gabungan).

2.3.4 Klasifikasi CNN

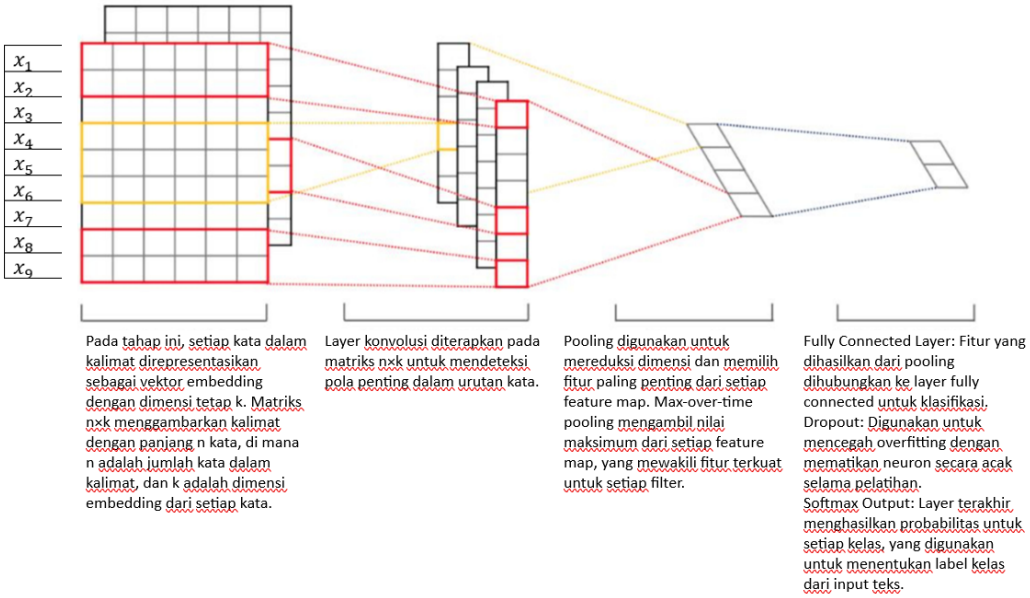
Klasifikasi teks mengotomatiskan kategorisasi teks ke dalam kelas yang telah ditentukan sebelumnya, menjadikan pengelolaan data lebih efisien. Manfaat klasifikasi antar lain sistem pengambilan teks yang merespons pertanyaan pengguna yang bertujuan memahami teks, seperti sistem peringkasan atau penjawab pertanyaan (Kamruzzaman et al., 2010), meningkatkan ketepatan sistem pengambilan informasi dengan memastikan hanya informasi relevan yang diambil sesuai dengan kebutuhan pengguna (Nithya et al., 2012), pemfilteran spam, identifikasi bahasa, analisis sentiment (Vijayan et al., 2017). Sistem klasifikasi teks modern dirancang agar dapat beradaptasi dengan berbagai jenis data dan cukup akurat untuk memenuhi harapan pengguna, sehingga memberikan manfaat bagi berbagai domain termasuk manajemen konten web dan penanganan data perusahaan (X. Zhou et al., 2020).

SEKOLAH PASCASARJANA

Penelitian ini mengusulkan algoritma CNN untuk klasifikasi karena pada beberapa penelitian menggunakan CNN menghasilkan akurasi yang baik yaitu mengklasifikasikan pesan singkat online dalam bahasa Mandarin dengan hasil efektif dalam evaluasi label (Lu et al., 2023), klasifikasi teks pendek berbasis CNN meningkatkan hasil klasifikasi teks pendek dengan menggabungkan model bahasa

N-gram dan mekanisme konsentrasi untuk pemilihan fitur (H. Wang et al., 2020). CNN untuk klasifikasi teks meningkatkan resistensi terhadap overfitting dan meningkatkan kinerja pada lima kumpulan data benchmark (L. Li et al., 2020).

Bagian teks dalam Gambar 2.2 mendeskripsikan arsitektur dasar CNN yang digunakan untuk penyisipan urutan dalam pemrosesan teks yang dirujuk dari (Fesseha et al., 2021).



Gambar 2.2 Tahapan Klasifikasi CNN (Fesseha et al., 2021)

Tahapan CNN untuk klasifikasi teks sebagai berikut:

1. Input Layer (Lapisan Input)

Dalam pemrosesan teks menggunakan model konvolusi, satu kalimat yang terdiri dari beberapa kata x1,x2,…,xn direpresentasikan sebagai konkatenasi penggabungan) dari vektor-vektor kata tersebut. Notasi ini dilambangkan dengan

$$x_{1:n}=x_1 \oplus x_2 \oplus \cdots \oplus x_n \quad (2.6)$$

Simbol $\oplus$ merepresentasikan operator konkatenasi, yaitu penggabungan vektor-vektor kata secara berurutan. Misalnya, jika menyebutkan $x_{i:i+j}$, ini berarti merujuk pada penggabungan vektor kata dari $x_i, x_{i+1}, \ldots . x_{i+j}$

2. Operasi Konvolusi:

Konvolusi adalah proses menerapkan filter atau kernel w dengan dimensi tertentu, yaitu R^{hk} (dengan h sebagai tinggi atau ukuran jendela filter), pada serangkaian kata dalam kalimat untuk menghasilkan fitur baru. Filter ini akan diaplikasikan pada jendela kata yang mencakup h kata secara bersamaan untuk menangkap pola atau hubungan tertentu antara kata-kata tersebut.

Sebagai contoh, sebuah fitur C_i dihasilkan dari jendela kata $x_{i:i+h-1}$ dengan rumus:

$$C_i = f(w \,.\, x_{i:i+h-1} + b) \quad (2.7)$$

$w \,.\, x_{i:i+h-1}$ adalah hasil perkalian (dot product) antara filter w dengan vektor kata yang sudah dikonkatenasi dalam jendela $x_{i:i+h-1}$. Ini merepresentasikan penggabungan informasi yang berasal dari jendela kata tersebut.

b ∈ R adalah istilah bias yang ditambahkan untuk menyesuaikan hasil konvolusi.

f adalah fungsi aktivasi non-linear Relu, yang diterapkan untuk memperkenalkan non-linearitas pada hasil. Non-linearitas ini penting agar model dapat menangkap pola yang kompleks.

3. Peta Fitur (Feature Map):

Filter konvolusi ini diterapkan pada setiap jendela kata dalam kalimat. Dengan cara ini, menghasilkan sebuah feature map yang merepresentasikan pola atau fitur relevan yang diambil dari seluruh jendela dalam urutan kata pada kalimat.

Misalnya, jika kalimat memiliki 10 kata dan filter memiliki ukuran jendela 3 (artinya h=3), maka filter tersebut akan diterapkan pada jendela-jendela bertumpuk: $x_{1:3}, x_{2:4}, x_{3:5}, \dots, x_{8:10}$ menghasilkan serangkaian fitur $c_1, c_2, c_3, \dots, c_8$ yang membentuk peta fitur.

4. Max Pooling

SEKOLAH PASCASARJANA

Setelah konvolusi menghasilkan peta fitur (feature map), max pooling diterapkan untuk mengurangi dimensi peta fitur dan memilih fitur yang paling menonjol. Max pooling adalah operasi yang memilih nilai maksimum dari peta fitur, yaitu:

$$\hat{C} = \max\{c_1, c_2, c_3, \dots, c_{n-h+1}\} \quad (2.8)$$

Tujuan max pooling adalah untuk menangkap fitur terpenting dari hasil konvolusi dan mereduksi dimensi data agar model menjadi lebih sederhana dan cepat. Dalam konteks teks, max pooling membantu model menangkap fitur paling relevan atau signifikan dalam suatu kalimat.

Max pooling mengatasi variasi panjang kalimat dengan hanya mengambil satu nilai maksimum, yang mewakili seluruh kalimat untuk filter.

Tahap Fully Connected Layer, Setelah pooling, proses menggabungkan nilai maksimum yang dihasilkan oleh setiap filter untuk membentuk vektor fitur yang akan diproses oleh lapisan fully connected. Misalkan menggunakan beberapa filter dengan ukuran jendela yang berbeda, maka hasil pooling dari setiap filter digabungkan menjadi vektor

$$z = [\hat{C}_1, \hat{C}_2, \dots, \hat{C}_m] \quad (2.9)$$

dengan m adalah jumlah filter yang digunakan.

Vektor z ini berisi fitur-fitur penting yang dideteksi oleh setiap filter dan mewakili representasi dari seluruh kalimat setelah diproses oleh konvolusi dan pooling. Lapisan fully connected menggunakan vektor z ini sebagai input, dan bobot lapisan fully connected akan belajar untuk mengoptimalkan kombinasi fitur ini untuk klasifikasi. Lapisan fully connected bertindak sebagai penghubung antara fitur-fitur yang dihasilkan oleh CNN dan lapisan output yang digunakan untuk klasifikasi.

5. Tahap Softmax untuk Klasifikasi

Pada tahap terakhir, lapisan fully connected terhubung dengan lapisan softmax untuk menghasilkan probabilitas untuk setiap kelas dalam tugas klasifikasi. Fungsi softmax mengubah output dari lapisan fully connected menjadi distribusi probabilitas, sehingga setiap nilai menunjukkan probabilitas kalimat tersebut termasuk dalam masing-masing kelas.

Rumus softmax untuk kelas ke-j adalah:

$$P(y = j|x) = \frac{e^{(W_{j.z} + b_j)}}{\sum_k e^{(W_{k.z} + b_k)}} \quad (2.10)$$

W_j adalah bobot untuk kelas j,

b_j adalah bias untuk kelas j,

SEKOLAH PASCASARJANA

z adalah vektor fitur yang dihasilkan dari pooling dan fully connected layer. Hasil dari fungsi softmax adalah probabilitas untuk setiap kelas, dan kelas dengan probabilitas tertinggi akan dipilih sebagai prediksi akhir model.

6. Training

Proses ini diulangi selama beberapa epoch, di mana satu epoch mencakup satu kali iterasi penuh atas seluruh dataset. Batch data dibagi menjadi batch, dan pembaruan bobot terjadi setelah memproses setiap batch (*mini-batch gradient descent*).

7. *Evaluation*

Evaluasi menggunakan *Accuracy, Precision, Recall*. Setelah pelatihan, model dievaluasi menggunakan matrik evaluasi seperti akurasi, *precision, recall, dan F1-score* untuk menilai kinerja model pada data uji.

Setelah model CNN dilatih, langkah pertama dalam evaluasi adalah menghasilkan prediksi untuk data uji yang belum pernah dilihat oleh model. Data uji ini harus memiliki label yang diketahui (ground truth) sehingga dapat dibandingkan prediksi model dengan label yang sebenarnya.

Probabilitas dan Label Prediksi: Untuk setiap sampel teks dalam data uji, model akan menghasilkan probabilitas untuk setiap kelas yang mungkin. Kelas dengan probabilitas tertinggi dipilih sebagai label prediksi akhir.

Confusion Matrix adalah tabel yang digunakan untuk menilai kinerja model klasifikasi. Tabel ini mengukur prediksi model versus label sebenarnya. Komponen Confusion Matrix yaitu: True Positive (TP): Jumlah sampel yang benar-benar positif dan diprediksi sebagai positif. True Negative (TN): Jumlah sampel yang benar-benar negatif dan diprediksi sebagai negatif. False Positive (FP): Jumlah sampel yang benar-benar negatif tetapi diprediksi sebagai positif. False Negative (FN): Jumlah sampel yang benar-benar positif tetapi diprediksi sebagai negatif.

Menghitung Akurasi:

Akurasi akan mengukur proporsi prediksi yang benar (baik positif maupun negatif) terhadap jumlah total sampel.

$$akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.11)$$

SEKOLAH PASCASARJANA

Akurasi memberikan gambaran umum tentang seberapa sering model membuat prediksi yang benar. Namun, akurasi bisa keliru jika dataset sangat tidak seimbang.

Menghitung Precision, Precision adalah proporsi prediksi positif yang benar terhadap semua prediksi positif yang dibuat oleh model.

$$Precision = \frac{TP}{TP+FP} \quad (2.12)$$

Precision menilai seberapa baik model menghindari kesalahan positif (false positives). Ini penting dalam konteks di mana kesalahan positif bisa sangat merugikan.

Menghitung Recall (Sensitivity). Recall adalah proporsi sampel positif yang benar-benar teridentifikasi oleh model sebagai positif.

$$Recall = \frac{TP}{TP+FN} \quad (2.13)$$

Recall mengukur kemampuan model untuk menemukan semua contoh positif dalam data. Ini penting dalam konteks di mana kesalahan negatif (false negatives) harus diminimalkan.

Menghitung F1-Score**:**

F1-Score: F1-Score adalah rata-rata harmonik dari precision dan recall. Ini memberikan keseimbangan antara keduanya dan lebih berguna daripada akurasi dalam konteks data tidak seimbang.

$$F1 - Score = \frac{PrecisionxRecall}{Precision+Recall} \quad (2.14)$$

F1-Score memberikan satu angka yang menilai keseimbangan antara kemampuan model untuk tidak mengabaikan sampel positif (recall) dan tidak salah mengidentifikasi sampel negatif sebagai positif (precision).

Tahapan CNN untuk klasifikasi data teks adalah sebagai berikut:

Input Layer: Inputan untuk CNN adalah data dalam bentuk matriks. Untuk data teks, input berupa matriks representasi teks Word Embedding.

$$X = [x_{1,1}, x_{1,2}, \dots, x_{m,n1}]$$

Dengan $x_{i,j}$ adalah elemen data input, m×n adalah dimensi matriks input.

Convolutional Layer bertujuan untuk mengekstraksi fitur dengan cara melibatkan filter (kernel) yang akan di konvolusi dengan matriks input.

SEKOLAH PASCASARJANA

Operasi Konvolusi:

$$z_{i,j} = \sum_{p=1}^{k} \sum_{q=1}^{k} W_{(p,q)} \cdot X_{i+p-1,j+q-1} + b$$

Dengan $z_{i,j}$adalah nilai hasil konvolusi di posisi (i,j), $w_{p,q}$ adalah bobot kernel/filter dengan ukuran k×k, B adalah bias, $X_{i+p-1,j+q-1}$adalah elemen input, Output dari layer ini disebut Feature Map.

Setelah konvolusi, dilakukan operasi aktivasi untuk memperkenalkan non-linearitas. Setelah tahapan aktivasi proses selanjutnya adalah ReLU Activation:

$$a_{i,j} = \max(0, z_{i,j})$$

Dengan $a_{i,j}$adalah hasil setelah aktivasi di posisi (i,j), $z_{i,j}$ adalah nilai dari operasi konvolusi.

Pooling layer digunakan untuk mengurangi dimensi feature map dan menjaga informasi penting.

Max Pooling: $p_{i,j} = \max(x_{i+p-1,j+q-1})$

Setelah pooling proses selanjutnya adalah Average Pooling

$$p_{i,j} = \frac{1}{k^2} \sum_{p=1}^{k} \sum_{q=1}^{k} x_{i+p-1,j+q-1}$$

Dengan: $p_{i,j}$ adalah nilai hasil pooling di posisi (i,j), k adalah ukuran filter pooling. p menunjukkan posisi baris relatif terhadap kernel pooling, q: menunjukkan posisi kolom relatif terhadap kernel pooling. Setelah beberapa lapisan konvolusi dan pooling, matriks di-flatten menjadi vektor 1D untuk masuk ke fully connected layer.

$$x = [x_1, x_2, \ldots, x_n]$$

dengan v_i adalah elemen vektor hasil flattening.

SEKOLAH PASCASARJANA

Setelah dilakukan flatten tahap selanjutnya adalah Fully Connected Layer (Dense Layer). Pada tahap ini, vektor hasil flattening dihubungkan dengan neuron-neuron dalam lapisan dense.

Operasi Neuron:

$$y_i = \sigma(\boldsymbol{w_i} \cdot \boldsymbol{x} + b_i)$$

Dengan y_i adalah output dari neuron i, $\boldsymbol{w_i}$ adalah bobot untuk neuron i, **x** adalah vektor input dari layer sebelumnya, b_i adalah bias untuk neuron i. σ adalah sigma fungsi aktivasi seperti softmax. Proses selanjutnya adalah Output Layer. Layer ini menghasilkan probabilitas untuk setiap kelas. Softmax Activation (untuk klasifikasi multi-kelas):

$$\hat{y}_j = \frac{e^{z_j}}{\sum_{k=1}^{n} e^{z_k}}$$

Dengan $\hat{y}_j$ adalah probabilitas kelas j, z_j adalah skor logit untuk kelas j. Skor logit digunakan untuk merepresentasikan kemungkinan relatif suatu input termasuk dalam kelas tertentu.

2.3.5 Data Tidak Seimbang

ROS merupakan teknik oversampling untuk menangani ketidakseimbangan kelas dengan cara meningkatkan jumlah sampel pada kelas minoritas. Teknik ini bekerja dengan cara mengambil sampel secara acak dari kelas minoritas dan menyalinnya hingga mencapai keseimbangan dengan kelas mayoritas. ROS menciptakan duplikat dari sampel yang sudah ada dalam kelas minoritas untuk meningkatkan representasi kelas tersebut dalam dataset.

NCL teknik penanganan data tidak seimbang, dengan menggabungkan konsep dari undersampling dan pembersihan data. NCL bekerja dengan memeriksa setiap sampel dalam kelas mayoritas dan menghapus sampel-sampel tersebut yang berada terlalu dekat (dalam artian karakteristik atau fitur) dengan kelas minoritas. Tujuannya adalah untuk membersihkan "neighborhood" dari kelas minoritas sehingga meminimalisir kesalahan klasifikasi yang disebabkan oleh sampel kelas mayoritas yang berdekatan dengan kelas minoritas

Mengatasi dataset yang tidak seimbang melibatkan berbagai teknik untuk mengurangi jumlah sampel pada kelas mayoritas (undersampling) atau meningkatkan jumlah sampel pada kelas minoritas (oversampling). Saat menggunakan kombinasi ROS (Kamalov et al., 2023; Johnson & Khoshgoftaar, 2020b) dan NCL (Al Abdouli et al., 2015) untuk data yang tidak seimbang, proses ini bertujuan untuk menyeimbangkan distribusi kelas dengan menggandakan

instance kelas minoritas secara acak dan menghapus instance tertentu dari kelas mayoritas berdasarkan dampaknya terhadap aturan nearest neighbor.

ROS bekerja dengan cara meningkatkan jumlah contoh kelas minoritas secara acak sampai jumlahnya mendekati atau sama dengan jumlah contoh kelas mayoritas, sehingga distribusi kelas dalam kumpulan data menjadi lebih seimbang. Teknik ROS dibutuhkan karena ketidakseimbangan kelas bisa berdampak negatif pada performa model pembelajaran mesin. Model cenderung bias terhadap kelas mayoritas dan menghasilkan prediksi yang kurang akurat untuk kelas minoritas. Dengan membuat distribusi kelas menjadi lebih seimbang, ROS membantu dalam meningkatkan akurasi dan keandalan model dalam mengklasifikasikan kelas minoritas.

Untuk menangani data yang tidak seimbang dengan menggunakan kombinasi ROS dan NCL dengan menambahkan lebih banyak instance dari kelas minoritas dan membersihkan instance dari kelas mayoritas yang kemungkinan merupakan noise. Tahapan detail dari pendekatan ini sebagai berikut:

Langkah 1 menggunakan algoritma ROS. Dalam dataset fitur Fintech memiliki 10 kelas, proses ROS dengan menyesuaikan sampel dari semua kelas minoritas sehingga jumlahnya sebanding dengan kelas yang memiliki jumlah sampel maksimum, dengan penjelasan sebagai berikut:

Dengan D adalah dataset yang terdiri dari K kelas, dengan K=10 Setiap kelas k memiliki n_k sampel, dimana k=1,2,...,10. Proses ini menyeimbangkan jumlah sampel di semua kelas sehingga semua kelas memiliki jumlah sampel yang mendekati sama dengan kelas yang memiliki jumlah sampel terbanyak. Menentukan kelas mayoritas dengan jumlah sampel maksimum $n_{max} =$ max $(n1, n2, \ldots n10)$. Menghitung Faktor pengali untuk setiap kelas untuk setiap kelas k, hitung faktor pengali f_k yang akan menentukan berapa kali setiap sampel dalam kelas tersebut harus diduplikasi:

$$f_k = \left\lceil \frac{n_{max}}{n_k} \right\rceil \tag{2.15}$$

Dengan $\lceil . \rceil$ adalah fungsi pembulatan ke atas.

SEKOLAH PASCASARJANA

Duplikasi Sampel: Untuk setiap sampel $x_{i,k}$ dari kelas k, duplikasi sampel tersebut sebanyak $f_k$kali dalam dataset yang telah diseimbangkan D'. Dataset yang dihasilkan setelah proses ini, jumlah sampel untuk setiap kelas k dalam dataset D' akan menjadi $n'_k = n_k \times f_k$ dan idealnya untuk setiap k, $n'_k \approx n_{max}$.

Notasi D adalah dataset awal dengan 10 kelas, D' adalah dataset setelah penerapan ROS, $n_k$adalah Jumlah sampel di kelas k sebelum ROS, n'_k: Jumlah sampel di kelas k setelah ROS, n_{max} adalah jumlah sampel maksimum dari kelas manapun dalam dataset D, f_k adalah faktor pengali untuk kelas k.

Langkah 2 menggunakan algoritma NCL. NCL menghapus instance dari kelas mayoritas yang dianggap sebagai noise atau kemungkinan menyebabkan misclassification dari sampel kelas minoritas. Teknik ini memanfaatkan pendekatan Nearest Neighbors untuk mengidentifikasi sampel seperti itu. Sebuah Tomek Link merupakan pasangan sampel (x_i, x_j) dari dua kelas berbeda dimana kedua sampel tersebut adalah nearest neighbors satu sama lain dan tidak ada sampel lain yang lebih dekat dengan salah satu dari mereka. Dalam kata lain, tidak ada sampel ketiga x_k sehinggad$(x_i, x_k) < d(x_i, x_j)$ atau $d(x_j, x_k) < d(x_i, x_j)$ dengan d menunjukkan suatu fungsi jarak (jarak Euclidean).

Setelah proses ROS dilanjutkan masuk ke proses NCL dengan jumlah dataset 10 kelas. Langkah sistematik untuk menerapkan NCL pada dataset fitur Fintech 10 kelas yaitu langkah 1 melakukan identifikasi kelas mayoritas dan minoritas. Proses pertama mengidentifikasi kelas yang merupakan kelas mayoritas (yang memiliki jumlah sampel terbanyak) dan kelas-kelas minoritas (yang memiliki jumlah sampel lebih sedikit). Ini dapat ditentukan dengan menghitung jumlah sampel per kelas.

Diketahui dataset D dengan sampel $\{x_1, x_2 \ldots\ldots x_n\}$ dan label kelas $\{y_1, y_2 \ldots\ldots y_n\}$ dengan $y_i \in \{1,2, \ldots,10\}$.

SEKOLAH PASCASARJANA

Langkah 2 menerapkan aplikasi ENN (Edited Nearest Neighbors) pada kelas mayoritas, untuk setiap sampel x_i dari kelas mayoritas, hitung k-nearest neighbors (menggunakan jarak Euclidean). Jika mayoritas dari k tetangga terdekat memiliki label yang berbeda dari xi, hapus xi dari dataset.

Formula Jarak Euclidean:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2} \quad (2.16)$$

Dengan m adalah jumlah dimensi.

Langkah 3 melakukan identifikasi dan hapus tomek links, untuk setiap pasangan sampel (x_i, x_j) dengan $x_i$dan x_j adalah tetangga terdekat dan berasal dari kelas yang berbeda. Jika tidak ada sampel yang lebih dekat ke x_i atau x_j satu sama lain, maka membentuk sebuah tomek link. Hapus sampel dari kelas mayoritas yang terlibat dalam Tomek Link untuk membantu mengklarifikasi batas antar kelas.

Langkah 4 mengulangi langkah 2 dan 3 sampai konvergen (tidak ada lagi sampel yang dapat dihapus), menunjukkan bahwa batas antar kelas telah dioptimalkan.

SEKOLAH PASCASARJANA