Scientax
Jurnal Kajian Ilmiah Perpajakan Indonesia

djp

VOLUME 7 NO. 2 | APRIL 2026 Journal Page: ejurnal.pajak.go.id ISSN 2686-5718

Penerapan Model *Natural Language Processing (LDA-BOW & Word2Vec)* & Diagram Sankey dalam Analisis Rantai Pasokan pada Bidang Perpajakan

Ryan Agatha Nanda Widiiswa[a*], Sigit Hariyanto[b], Muhammad Rifqi Aziz [c] Ajar Parama Adhi[d]
[a] Direktorat Jenderal Pajak, Jakarta, Indonesia. Email: ryan.widiiswa@pajak.go.id
[b] Direktorat Jenderal Pajak, Jakarta, Indonesia. Email: sigit.hariyanto@pajak.go.id
[c] Direktorat Jenderal Pajak, Jakarta, Indonesia. Email: rifqi.aziz@pajak.go.id
[d] Direktorat Jenderal Pajak, Jakarta, Indonesia. Email: ajar.adhi@pajak.go.id

*Penulis korespondensi: ryan.widiiswa@pajak.go.id

ABSTRACT

The development of digital innovation and big data has created a growing need for the application of machine learning across various fields. This study seeks to integrate existing tax analysis practices with machine learning techniques. It aims to improve the efficiency and accuracy of supply chain analysis in the tax domain by applying Natural Language Processing (NLP) models alongside Sankey diagram visualizations. The NLP models employed include Latent Dirichlet Allocation Bag of Words (LDA BOW) and the Word2Vec algorithm, which serve to identify and extract transactions based on topic modeling and semantic similarity. These models are implemented within the CRISP-DM methodological framework. As a result of this application, 6.8 million PKP transactions in the pharmaceutical sector for the year 2022 were successfully classified at a rate of 73.7 %, with a 19 % improvement in accuracy following the integration of Word2Vec. In this research, Sankey diagrams are used to intuitively visualize the flow of transactions, enabling users to pinpoint critical points in the supply chain where tax-related risks or discrepancies are higher. For the supply chain analysis, the authors adopt the Supply Chain Operations Reference (SCOR) model, focusing on reliability and cost aspects that closely align with tax compliance evaluation. The findings are expected to yield a prototype application that streamlines the audit process for tax authorities and contributes to the body of text-mining literature in the field of taxation.

Keywords: natural language processing, supply chain analysis, machine learning, Sankey diagram, tax analysis

ABSTRAK

Perkembangan era inovasi digital serta *big data* menumbuhkan kebutuhan penerapan *machine learning* pada berbagai bidang. Penelitian ini mencoba untuk memadukan kegiatan analisis perpajakan yang ada dengan *machine learning*. Penelitian ini mencoba meningkatkan efisiensi dan akurasi pada analisis rantai pasokan di bidang perpajakan dengan menerapkan model *Natural Language Processing* (NLP) dan visualisasi diagram Sankey. Model NLP yang digunakan meliputi *Latent Dirichlet Allocation Bag of Words* (LDA BOW) dan algoritma *Word2Vec*, yang berfungsi untuk mengidentifikasi dan mengekstraksi transaksi berdasarkan topik dan kesamaan semantik. Model tersebut

DOI: 10.52869/ad83d368
Diunggah: 19 November 2024; Direvisi: 29 Juni 2025; Diterima: 27 Maret 2026; Diterbitkan: 30 April 2026

Cara Mengutip:
Widiisma, R. A. N., Hariyanto, S., Aziz, M. R., & Adhi, A. P. (2026). Penerapan model natural language processing (LDA-BOW & Word2Vec) & diagram Sankey dalam analisis rantai pasokan pada bidang perpajakan. *Scientax: Jurnal Kajian Ilmiah Perpajakan Indonesia, 7*(2), 248-270. https://doi.org/10.52869/ad83d368

diterapkan dalam kerangka metode *CRISP-DM*. Dengan hasil penerapan, 6,8 juta transaksi PKP sektor farmasi tahun 2022, telah berhasil diklasifikasikan sebesar 73,7%, dengan peningkatan akurasi sebesar 19% setelah integrasi *Word2Vec*. Diagram Sankey pada penelitian ini digunakan untuk memvisualisasikan aliran transaksi secara intuitif, serta memungkinkan pengguna mengidentifikasi titik-titik kritis dalam rantai pasokan yang mengalami risiko atau ketidaksesuaian lebih tinggi dari aspek perpajakan. Dalam melakukan analisis rantai pasokan, penulis menggunakan model *Supply Chain Operation Reference (SCOR)* dengan fokus pada aspek keandalan dan biaya, yang dekat dengan analisis kepatuhan perpajakan. Hasil penelitian diharapkan dapat menyediakan prototipe aplikasi yang mempermudah fiskus dalam proses audit dan memperkaya literatur *text-mining* di bidang perpajakan.

Kata kunci: natural language processing, analisis rantai pasokan, machine learning, diagram Sankey, analisis pajak

1. PENDAHULUAN

Seiring dengan kemajuan teknologi yang merombak berbagai industri, integrasi antara bidang perpajakan dengan *machine learning* menjadi hal yang tidak dapat dihindarkan dewasa ini. Integrasi tersebut dapat meningkatkan efektivitas dan efisiensi dalam berbagai kegiatan perpajakan (Alarie et al., 2018). *Organisation for Economic Co-operation and Development* (OECD) dalam laporannya pada tahun 2021 menyebutkan menekankan bahwa pemanfaatan teknologi canggih, termasuk *machine learning*, adalah kunci untuk modernisasi administrasi pajak yang lebih responsif adaptif terhadap tantangan masa depan. OECD memberikan contoh pada otoritas pajak Inggris, HM *Revenue and Customs* (HMRC) memanfaatkan *machine learning* untuk memproses dan menganalisis data dari berbagai sumber untuk mengungkap potensi skema penghindaran pajak

Di Indonesia, analisis rantai pasokan merupakan salah satu aktivitas yang paling potensial untuk diautomasi karena kompleksitas data dan tingginya volume transaksi yang selama ini diperiksa secara manual. Secara konseptual, rantai pasokan melacak aliran barang, informasi, dan dana mulai bahan baku hingga produk mencapai konsumen (Hugos, 2018). Bagi korporasi, analisis ini membantu menekan inefisiensi biaya dan mengelola risiko operasional (Sürie & Wagner, 2005).

Dari sudut pandang institusi perpajakan sendiri, analisis rantai pasokan dapat digunakan untuk membantu investigasi penerapan strategi *tax avoidance* wajib pajak dari sudut pandang transaksi antar pemasok dan pelanggan (Cen et al., 2016). Analis rantai pasokan dapat digunakan untuk membantu fiskus melakukan pengujian ketidaksesuaian bahan baku yang dibeli dan pengeluaran yang dikeluarkan dengan barang jadi yang diperjualbelikan oleh sebuah perusahaan. Selain itu, analisis rantai pasokan juga dapat dimanfaatkan oleh fiskus untuk mengidentifikasi harga produk jual beli antar entitas sesuai dengan harga pasar, sehingga dapat mengetahui risiko manipulasi harga untuk tujuan penghindaran pajak. Fiskus juga dapat melakukan deteksi adanya barang produksi yang tidak dilaporkan melalui analisis rantai pasokan (Cen et al., 2016; OECD, 2021). Integrasi analisis rantai pasokan dalam bidang perpajakan berbasis *machine learning* merupakan pendekatan yang menjanjikan. Relevansinya semakin kuat mengingat kewajiban hukum yang melekat pada Pengusaha Kena Pajak (PKP), yakni pencatatan setiap transaksi melalui faktur pajak (berdasarkan Undang-Undang Nomor 8 Tahun 1983) serta pelaporan dokumen Pemberitahuan Impor Barang (PIB) dan Pemberitahuan Ekspor Barang (PEB) untuk setiap transaksi lintas batas (berdasarkan Undang-Undang Nomor 10 Tahun 1995). Kewajiban pencatatan tersebut menghasilkan rekam jejak transaksi yang kaya dan terstruktur, sehingga berpotensi menjadi sumber data yang andal bagi pengembangan model *machine learning* dalam konteks pengawasan perpajakan. Sayangnya, deskripsi barang pada faktur, PIB, dan PEB bersifat tidak baku. Selain itu, PER-11/PJ/2023 memperkenalkan e-Faktur 3.0 dan PER-03/PJ/2024 mewajibkan standarisasi penulisan deskripsi barang/jasa dengan kode harmonisasi. Meskipun begitu, praktik di lapangan menunjukkan banyak WP masih menuliskan uraian barang secara tidak baku, sehingga Direktorat

Jenderal Pajak masih harus menangani sekitar 3,4 miliar baris faktur pada tahun 2023 (angka tersebut tumbuh konsisten 9–10 % per tahun).

Berangkat dari kondisi tersebut, penulis merasa bahwa integrasi penerapan *machine learning* dengan analisis rantai pasokan menjadi hal yang tidak terhindarkan. Bagaimana nantinya fiskus ataupun pembuat kebijakan dapat melakukan analisis rantai pasokan secara efektif dan efisien dengan jumlah data jutaan bahkan ratusan juta dalam waktu hitungan menit. Model *machine learning* yang diterapkan penulis adalah teknik *Natural Language Processing* (NLP) serta visualisasi diagram Sankey. Teknik NLP yang digunakan untuk memetakan topik, kata, serta konteks dari isian pada faktur pajak dan PIB-PEB adalah algoritma *Latent Dirichlet Allocation* (LDA) *Bag of Words* (BOW). Pendekatan LDA BOW nantinya juga dapat menyimpan detail dari topik transaksi perusahaan seperti merek obat, merek bahan kimia, hingga jenis jasa. Topik-topik dari transaksi tersebut disempurnakan menggunakan algoritma *Word2Vec*. Ekstraksi topik atas setiap transaksi perusahaan memiliki peranan penting dalam konteks analisis potensi penyimpangan dan risiko pajak perusahaan. Dengan mengklasifikasikan setiap transaksi dengan jelas, institusi perpajakan dapat dengan mudah melacak aliran barang dan jasa serta mendeteksi pola yang tidak biasa atau mencurigakan yang dapat mengindikasikan adanya penghindaran pajak.

Hasil dari pemetaan topik tersebut divisualisasi melalui diagram Sankey. Kelebihan dari diagram Sankey sendiri adalah fitur visualisasi atas aliran dan transisi dari sebuah kejadian (Otto et al., 2022). Diagram Sankey memungkinkan pengguna untuk melihat aliran barang dari satu perusahaan ke perusahaan lain. Dengan visualisasi ini, fiskus dapat dengan cepat mengidentifikasi titik-titik kritis dan anomali dalam aliran transaksi, sehingga meningkatkan efisiensi dan akurasi dalam proses pengawasan dan penegakan hukum perpajakan. Lebih lanjut, analisis yang tepat dapat membantu organisasi perpajakan untuk mengidentifikasi area-area yang memerlukan audit lebih lanjut. Kegiatan analisis perpajakan tersebut dilaksanakan melalui analisis rantai pasokan. Analisis rantai pasokan yang penulis gunakan pada penelitian ini adalah analisis rantai pasokan model *Supply Chain Operation Reference* (SCOR) dengan sudut pandang *reliability* serta *cost*. Sudut pandang *realibility* melihat kesesuaian transaksi pembelian perusahaan dengan penjualan perusahaan, sedangkan sudut pandang *cost* melihat jenis pengeluaran perusahaan. Penelitian ini menggunakan data sumber faktur pajak, PIB, serta PEB Wajib Pajak yang telah berstatus PKP dibidang farmasi rentang tahun pajak 2022 dengan total populasi 6.897.594 dokumen transaksi. Data penelitian ini telah dilakukan *masking* identitas perusahaan. Hingga studi ini dibuat belum ada studi Indonesia yang menggabungkan LDA-BOW, Word2Vec, CRISP-DM, dan diagram Sankey untuk analisis rantai pasokan berbasis teks dari dokumen perpajakan. Penelitian ini berusaha mengisi celah tersebut serta menguji efektivitas pendekatan NLP pada data perpajakan berskala jutaan transaksi.

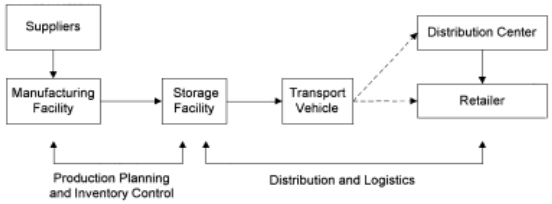
2. KERANGKA TEORETIS DAN PENGEMBANGAN HIPOTESIS

2.1 Analisis Rantai Pasokan

Proses rantai pasokan adalah sebuah proses terintegrasi antara pemasok, produsen, distributor, dan pengecer yang saling bekerja sama untuk: (1) memperoleh bahan baku, (2) mengubah bahan baku ini menjadi produk akhir yang spesifik, dan (3) mengirimkan produk akhir ini kepada *retailer* akhir (Beamon, 1998). Mekanisme alur rantai pasokan tersebut secara garis besar dapat dicermati pada diagram yang disajikan dalam Gambar 1.

Pada Gambar 1 dapat dilihat bahwa tahapan dari rantai pasokan dapat dipisah menjadi lima tahapan, dimulai dari bahan baku yang dibeli dari *supplier*, kemudian dilakukan pengolahan

Gambar 1
Alur Rantai Pasokan



Catatan. Sumber: Beamon (1998)

pada pabrik, hasil olahan kemudian ditaruh pada gudang, hingga akhirnya diserahkan pada *retailer* ataupun distributor (Beamon, 1998). Sedangkan analisis rantai pasokan adalah proses evaluasi menyeluruh dari setiap elemen dan aktivitas dalam rantai pasokan untuk meningkatkan efisiensi, mengurangi biaya, dan meningkatkan kinerja secara keseluruhan. Dalam melaksanakan analisis rantai pasokan, terdapat beberapa model yang tersedia, salah satunya, *Supply Chain Operation Reference (SCOR)*. SCOR model memberikan panduan untuk mengukur kinerja rantai pasokan melalui lima proses utama: *Plan* (perencanaan), *Source* (pengadaan), *Make* (produksi), *Deliver* (pengiriman), dan *Return* (pengembalian). Dengan menggunakan SCOR, perusahaan dapat mengevaluasi dan membandingkan kinerja merka terhadap praktik terbaik industri (Supply Chain Council, 2012). SCOR sendiri memiliki terdapat beberapa atribut kinerja yang terdiri atas (Supply Chain Council, 2012):

a. *Supply Chain Reliability*
Terkait kinerja rantai pasokan dalam mengirimkan produk yang sesuai ke lokasi yang tepat, dalam kondisi dan kemasan yang benar, dalam jumlah yang tepat, dengan dokumentasi yang benar, serta kepada konsumen yang tepat. Kinerja ini juga melihat bagaimana produksi barang yang dijual telah sesuai dengan bahan baku yang diolah mulai dari kuantitas hingga kualitas.

b. *Supply Chain Responsiveness*
Kecepatan rantai pasokan dalam memenuhi pesanan konsumen.

c. *Supply Chain Agility*
Agilitas rantai pasokan dalam merespons perubahan pasar.

d. *Supply Chain Cost*
Biaya yang berkaitan dengan rantai pasokan (*Cost of Good Sold*) apakah telah berjalan secara efektif dan efisien dan sesuai dengan peruntukan proses bisnis perusahaan.

e. *Supply Chain Asset Management*
Efektivitas pengelolaan manajemen asset untuk memenuhi permintaan pasar.

Penelitian ini berfokus pada sudut pandang keandalan rantai pasok (*supply chain reliability*) dan biaya rantai pasok (*supply chain cost*). Pemilihan kedua aspek tersebut didasarkan pada relevansinya yang erat dengan analisis kepatuhan perpajakan (Supply Chain Council, 2012).

Dalam konteks ini, *supply chain reliability* ditinjau dari kesesuaian antara produk yang dipasarkan dengan bahan baku yang digunakan oleh perusahaan. Sementara itu, *supply chain cost* berkaitan dengan beban dan biaya yang dikeluarkan perusahaan atas produk yang dijual.

2.2 Dokumen Transaksi Perusahaan

Pada Undang-Undang Nomor 8 Tahun 1983 tentang Pajak Pertambahan Nilai Barang dan Jasa dan Pajak Penjualan atas Barang Mewah (UU PPN) Pasal 3 menyatakan bahwa setiap pengusaha yang menyerahkan Barang Kena Pajak (BKP) dan/atau Jasa Kena Pajak (JKP) wajib melaporkan usahanya untuk dikukuhkan sebagai Pengusaha Kena Pajak (PKP). Dari sisi kewajiban yang harus dipenuhi PKP, jika dilihat pada pasal 13 UU PPN menyebutkan bahwa PKP harus membuat faktur pajak atas setiap penyerahan BKP dan/atau JKP. Lebih lanjut, pembuatan faktur harus dibuat oleh PKP pada momen:

a. penyerahan BKP di dalam Daerah Pabean yang dilakukan oleh PKP;
b. penyerahan JKP di dalam Daerah Pabean yang dilakukan oleh PKP;
c. ekspor BKP tidak berwujud, dan/atau ekspor JKP oleh PKP.

Sementara itu, untuk ekspor BKP berwujud dan impor BKP berwujud, setiap perusahaan tidak terbatas hanya pada perusahaan PKP, wajib membuat dokumen Pemberitahuan Impor Barang (PIB) untuk impor BKP berwujud dan Pemberitahuan Ekspor Barang (PEB) untuk ekspor BKP berwujud. Sehingga dapat dikatakan melalui dokumen faktur pajak serta PIB-PEB, penulis dapat mengetahui seluruh transaksi perusahaan yang telah berstatus PKP. Selanjutnya terkait dengan elemen informasi yang terdapat pada faktur pajak berdasarkan Peraturan Direktur Jenderal Pajak Nomor PER-24/PJ/2012 tentang Bentuk, Ukuran, Prosedur Pembuatan, Tata Cara Pembetulan atau Penggantian, dan Tata Cara Pembatalan Faktur Pajak pada Pasal 14, terdiri atas:

a. nama, alamat, dan NPWP yang menyerahkan BKP atau JKP;
b. nama, alamat, dan NPWP pembeli BKP atau penerima JKP;
c. jenis barang atau jasa, jumlah Harga Jual atau Penggantian, dan potongan harga;
d. PPN yang dipungut;
e. PPnBM yang dipungut;
f. kode, nomor seri, dan tanggal pembuatan faktur pajak; dan
g. nama dan tanda tangan yang berhak menandatangani faktur pajak.

Berpedoman pada Peraturan Menteri Keuangan Nomor 190/PMK.04/2022 tentang Pengeluaran Barang Impor untuk Dipakai serta Peraturan Menteri Keuangan Nomor 155/PMK.04/2022 tentang Ketentuan Kepabeanan di Bidang Ekspor, elemen utama PIB serta PEB adalah:

a. data identitas importir/eksportir;
b. data pengangkutan;
c. data barang;
d. data nilai;
e. data kepabeanan;
f. data pembayaran; dan
g. data pendukung.

Maka dari itu, dapat disimpulkan melalui dokumen faktur pajak, PIB, serta PEB; penulis dapat mengambil gambaran transaksi perusahaan yang terdiri mulai dari:

a. data *supplier*;
b. data *customer*;
c. nilai transaksi;
d. nama barang yang ditransaksikan; dan
e. masa dilakukan transaksi

untuk diolah dalam kebutuhan analisis rantai pasokan.

2.3 Hubungan Istimewa Wajib Pajak dan Penerapan dalam Analisis Rantai Pasokan

Hubungan istimewa merujuk pada hubungan antara dua atau lebih pihak yang dapat mempengaruhi harga atau syarat-syarat transaksi yang dilakukan di antara mereka, sehingga terjadi perbedaan perlakuan dengan pihak independen lainnya (Park, 2018). Berdasarkan Undang-Undang Nomor 36 Tahun 2008 tentang Pajak Penghasilan (PPh) Pasal 18 ayat 4 disebutkan bahwa hubungan istimewa dapat terjadi apabila:

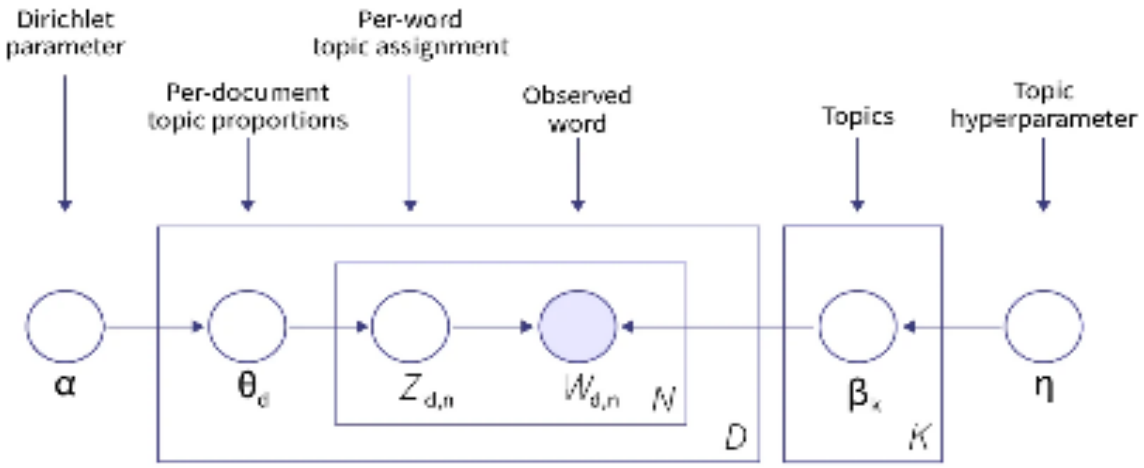
a. Kepemilikan
   Satu pihak memiliki penyertaan langsung atau tidak langsung sebesar 25% atau lebih pada pihak lain. Penyertaan ini bisa berupa saham atau ekuitas yang menunjukkan pengendalian atau kepemilikan signifikan.
b. Penguasaan
   Satu pihak menguasai pihak lain atau dua atau lebih pihak berada di bawah penguasaan yang sama. Penguasaan ini dapat berbentuk pengendalian manajemen, pengendalian operasional, atau pengendalian keuangan.
c. Hubungan Keluarga
   Terdapat hubungan keluarga baik sedarah maupun semenda dalam garis keturunan lurus satu derajat (misalnya, orang tua dan anak) atau garis ke samping satu derajat (misalnya, saudara kandung).

Dalam konteks analisis rantai pasokan, status hubungan istimewa wajib pajak memegang peranan penting. Hubungan ini dapat menimbulkan potensi manipulasi harga transfer atau *transfer pricing*, sehingga perusahaan dapat mengatur harga transaksi antarperusahaan untuk meminimalkan beban pajak secara keseluruhan. Praktik ini dapat mengurangi jumlah pajak yang seharusnya dibebankan oleh negara kepada perusahaan. Oleh karena itu, fiskus menerapkan prinsip kewajaran dan kelaziman usaha (*arm's length principle*) untuk memastikan bahwa harga dalam transaksi antara pihak-pihak yang memiliki hubungan istimewa dilakukan sebagaimana antara pihak-pihak yang tidak memiliki hubungan istimewa. Dalam penelitian ini, data status hubungan istimewa perusahaan diperoleh dari SPT Tahunan perusahaan.

2.4 Pengertian Machine *Learning*

Machine Learning dapat diartikan sebagai model komputasi yang memanfaatkan pengalaman untuk meningkatkan kinerja atau membuat prediksi yang akurat (Rostamizadeh et al., 2018). Bishop (2006)

Gambar 2
Alur Kerja LDA



Catatan. Sumber: Madala (2023)

dalam bukunya menyebutkan bahwa *machine learning* berkaitan dengan pengembangan algoritma dan model yang memungkinkan komputer untuk belajar. Hal ini memungkinkan komputer untuk melakukan tugas-tugas tanpa instruksi eksplisit, melainkan melalui pembelajaran dari data.

2.5 *Natural Language Process* dan Dokumen Transaksi Wajib Pajak

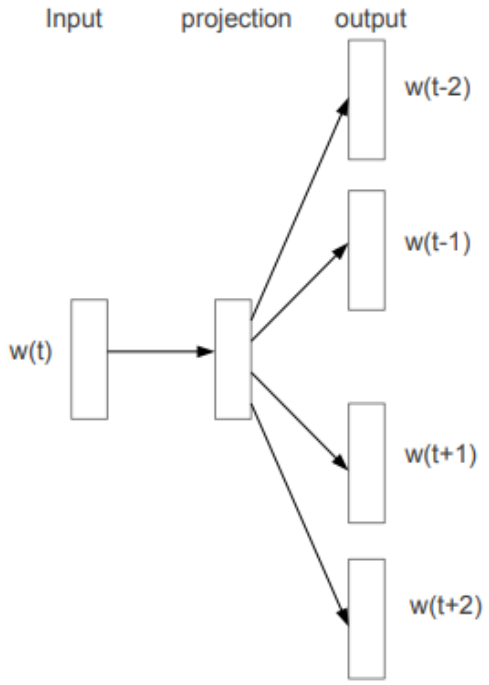
Natural language process (NLP) adalah bagian dari *machine learning* yang berisi kumpulan teknik yang bertujuan untuk menganalisis dan memodelkan bahasa manusia dengan bantuan sistem komputer (Manning et al., 2008). NLP secara garis besar memungkinkan komputer untuk memahami, menafsirkan, dan mengambil kesimpulan dari bahasa yang digunakan manusia secara alami. Lebih lanjut, NLP dapat digunakan untuk mengidentifikasi makna dan konteks pada teks yang bersifat tidak baku (Jurafsky & Martin, 2000).

Penelitian ini menggunakan algoritma *Latent Dirichlet Allocation* (LDA) *Bag of Words* (Bow) untuk memetakan topik dan merek dari isian pada faktur pajak dan PIB-PEB. Penelitian yang dilakukan oleh Chakkawar dan Tamane (2020) menyebutkan bahwa pendekatan LDA dengan fitur BOW merupakan pendekatan yang efektif dalam menentukan topik serta konteks dari sebuah dokumen tidak baku. LDA memodelkan topik dari sebuah dokumen, sehingga setiap topik dicirikan oleh distribusi kata-kata dari sebuah dokumen. Model ini memberikan representasi eksplisit dari sebuah dokumen dalam bentuk probabilitas topik (Blei et al., 2003). Dalam konteks ekstraksi topik dari sebuah teks, probabilitas topik memberikan representasi eksplisit dari sebuah dokumen. Kelebihan LDA BOW adalah tetap menyimpan tidak hanya topik dari transaksi perusahaan, tetapi juga detail dari jenis transaksi yang terjadi (misal obat ditulis detail menjadi obat sakit kepala mantab oye).

Alur kerja LDA dimulai dengan membuat topik yang diekstrak dalam sebuah dokumen melalui *text-corpus* (Madala, 2023). Penulis menyebutkan *text-corpus* dengan sebutan *Bag of Words*. Dengan kata lain, dokumen transaksi perusahaan yang nama barang dibeli tertera jenis merek obat, maka penulis memberikan topik Obat (contoh: Pembelian Obat Sakit Kepala XYZ sebanyak 1 dus, topik: Obat). Alur kerja LDA yang penulis sadur dari Madala (2023) tertera pada Gambar 2 alur kerja LDA.

Selain LDA BOW, penulis juga menerapkan algoritma *Word2Vec* pada hasil pemetaan LDA BOW. *Word2Vec* sendiri merupakan sebuah

Gambar 3
Arsitektur *Word2Vec* model *Skip-gram*



Catatan. Sumber: Mikolov et al. (2013)

algoritma yang menghasilkan representasi vektor untuk kata-kata dalam ruang vektor. Representasi ini memungkinkan untuk menggambarkan kata-kata yang memiliki makna serupa berada dekat satu sama lain dalam ruang vektor (Mikolov et al., 2013). Secara teknis, algoritma *Word2Vec* merepresentasikan susunan kata dalam sebuah kalimat ke dalam bentuk vektor, sehingga kata-kata dengan konteks korpus yang serupa akan berada pada ruang vektor yang berdekatan. Sebagai contoh, kata "paracetamol" akan memiliki nilai vektor yang mendekati kata "ibuprofen". Hal ini dikarenakan distribusi kemunculan kata "paracetamol" dalam sebuah kalimat memiliki kemiripan pola dengan kata "ibuprofen" jika dibandingkan dengan kata "laptop lenovo". Dalam penelitian ini, penulis menerapkan algoritma *Word2Vec* menggunakan model *skip-gram* dengan arsitektur sebagaimana ditampilkan pada Gambar 3.

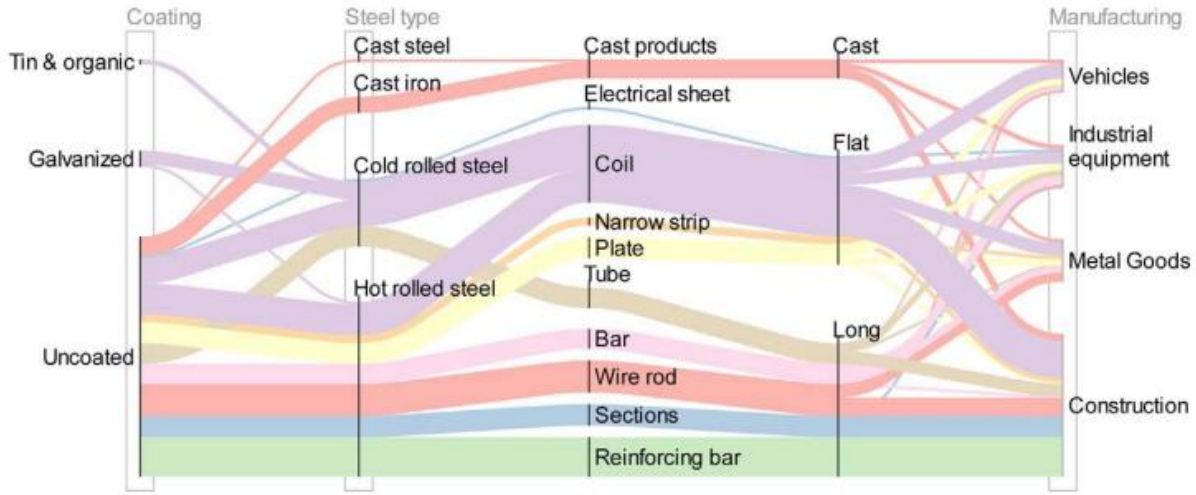
Penerapan *Word2Vec* pada penelitian ini didasari pada keterbatasan output dari LDA BOW. Dalam hal ini, algoritma LDA BOW mengabaikan semantik dan konteks dari sebuah kalimat, sehingga tidak dapat menangkap hubungan kata dalam kalimat. Algoritma *Word2Vec* diterapkan untuk mengatasi kelemahan tersebut (Turney & Pantel, 2010). Penerapan algoritma *Word2Vec* diharapkan dapat menangkap hubungan kontekstual semantik antarkata, sehingga dapat membantu kategorisasi jenis transaksi dalam analisis rantai pasokan. Artinya, melalui *Word2Vec*, penulis dapat melihat kedekatan sebuah kata yang memiliki topik serupa sekaligus memperluas cakupan *keywords* pada *Bag of Words*. Melihat konteks dari sifat uraian nama barang pada faktur pajak, PIB, dan PEB yang bersifat tidak baku serta memperhatikan fitur dari LDA BOW serta *Word2Vec*, penulis menggunakan LDA BOW serta *Word2Vec* untuk memetakan topik dari seluruh transaksi perusahaan. Oleh karena itu, seluruh transaksi perusahaan dapat diketahui topik serta jenis transaksinya untuk kebutuhan analisis rantai pasokan.

2.6 Diagram Sankey dan Visualisasi Transaksi

Diagram Sankey adalah jenis grafik yang menggambarkan aliran variabel antara berbagai tahap atau entitas. Lebar panah atau jalur dalam diagram ini sebanding dengan nilai aliran, sehingga memudahkan visualisasi yang jelas mengenai proporsi dan distribusi aliran atau dalam suatu sistem. Diagram Sankey dapat memvisualisasikan perubahan serta transisi yang terjadi pada setiap tahapan yang ingin diceritakan atas data yang divisualisasikan. Pada penelitian yang dilakukan oleh Lupton dan Allwood (2017) menyebutkan bahwa diagram Sankey dapat dimodifikasi sedemikian rupa untuk dapat menampilkan informasi dan hubungan pada multidimensional data, sebagaimana tergambar pada gambar 4 yang menggambarkan *Cullen's global steel flow data*.

Penulis berpendapat bahwa visualisasi menggunakan diagram Sankey atas transaksi faktur pajak yang telah ditentukan topiknya akan sangat bermanfaat bagi petugas pajak dalam menganalisis aliran transaksi dan memastikan kepatuhan terhadap prinsip kewajaran harga. Diagram Sankey memudahkan visualisasi aliran transaksi dan menunjukkan hubungan antar

Gambar 4
Diagram Sankey



Catatan. Sumber: Lupton dan Allwood (2017)

entitas dalam rantai pasokan secara intuitif, sehingga membantu petugas pajak mengidentifikasi titik-titik kritis dan anomali dalam aliran pajak serta meningkatkan efisiensi dan akurasi dalam proses pengawasan dan penegakan hukum.

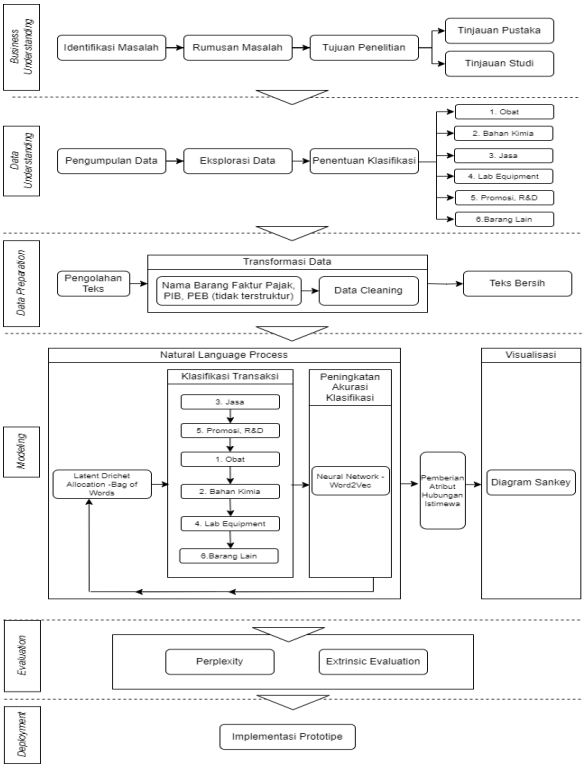
3. METODE PENELITIAN

Populasi pada penelitian ini adalah 6.897.594 dokumen transaksi Wajib Pajak (faktur pajak, PIB, & PEB) yang telah berstatus PKP di bidang industri farmasi pada tahun pajak 2022.

Penelitian ini memanfaatkan kerangka CRISP-DM untuk mengatur tahapan pengembangan prototipe, mulai dari *business understanding* hingga *deployment*. Model NLP diterapkan pada tahapan *modelling*. Pada tahap ini, *LDA Bag of Words* diperuntukkan untuk identifikasi barang jasa pada faktur pajak. Sementara itu, *Word2Vec* dimanfaatkan untuk ekspansi kosakata hasil klasifikasi transaksi. Pada fase *deployment*, hasil klasifikasi transaksi divisualisasikan sebagai diagram Sankey guna memudahkan analisis rantai pasokan.

Gambar 5 menunjukkan seluruh tahapan penelitian, mulai dari fase *business understanding* hingga proses *deployment* prototipe yang dikembangkan dalam beberapa lingkungan (*environment*). Dalam implementasinya, pengolahan data dilakukan menggunakan bahasa pemrograman Python, kemudian hasilnya diteruskan ke bagian *backend* yang menggunakan Django. *Output* dari *backend* tersebut berupa berkas JSON yang

Gambar 5
Tahapan Penelitian



Catatan. Sumber: Diolah penulis

kemudian dibaca oleh *frontend* berbasis JavaScript untuk divisualisasikan dalam diagram Sankey.

3.1 *Business Understanding*

Tahapan ini mencakup identifikasi masalah, baik yang terkait dengan bisnis maupun bersifat teknis. Dari sisi proses bisnis, institusi perpajakan memiliki kewajiban untuk melakukan pengujian kepatuhan perpajakan untuk memastikan Wajib Pajak mematuhi peraturan pajak. Hal ini sejalan dengan pilar-pilar kepatuhan poin penegakan yang telah ditetapkan oleh OECD (2004). Dalam konteks melaksanakan pengujian kepatuhan perpajakan dalam cakupan analisis rantai pasokan, institusi perpajakan dihadapkan dengan jumlah data yang cukup besar dari faktur pajak serta kompleksitas data yang ada. Dari kondisi tersebut, lahir urgensi adanya prototipe *machine learning* untuk menyelesaikan gap tersebut. Hal ini sejalan dengan laporan OECD tahun 2021 menyebutkan menekankan bahwa pemanfaatan teknologi canggih, termasuk *machine learning*, adalah kunci untuk pelaksanaan kegiatan perpajakan secara efektif dan efisien.

Dari sudut pandang teknis pembangunan prototipe sendiri, *machine learning* yang dipilih oleh penulis adalah model NLP *LDA BOW* & *Word2Vec*. Model ini dipilih untuk dapat mengklasifikasikan jenis transaksi perusahaan menjadi beberapa jenis kategori. Sifat dari isian nama barang pada dokumen faktur yang tidak baku, membuat pemilihan model NLP tersebut dipilih. LDA BOW & *Word2Vec* sangat efektif untuk memetakan topik dari dokumen yang tidak memiliki kalimat baku (Chakkarwar & Tamane, 2020). Model NLP LDA BOW digunakan untuk memetakan topik pada transaksi yang ada, sedangkan model *Word2Vec* digunakan untuk menjaring transaksi yang belum terpetakan melalui LDA BOW, termasuk memperluas cakupan *keywords* pada BOW.

Dalam memetakan topik dari dokumen transaksi, penulis mengklasifikasikan transaksi menjadi beberapa topik yang relevan dengan kegiatan proses bisnis perusahaan di bidang farmasi. Topik tersebut diharapkan dapat memecahkan masalah dalam analisis rantai pasokan dari sisi *realibility* dan *cost*. Transaksi yang telah dipetakan topiknya divisualisasikan dalam diagram Sankey. Visualisasi ini diharapkan dapat membantu pengguna maupun pengambil kebijakan dalam melakukan analisis rantai pasok. Melalui diagram tersebut, pengguna dapat menilai kesesuaian antara *input* dan *output* proses bisnis dengan klasifikasi usaha perusahaan (analisis *reliability* rantai pasok). Selain itu, pengguna dapat menguji kesesuaian jenis biaya dan jasa terhadap proses bisnis perusahaan (analisis *cost* rantai pasok).

3.2 *Data Understanding*

Tahap ini melibatkan eksplorasi data untuk memahami karakteristik dan struktur data yang tersedia. Data yang digunakan dalam penelitian ini adalah faktur pajak, PIB, dan PEB dari Wajib Pajak yang telah berstatus PKP di bidang farmasi selama rentang tahun 2018-2020. Data yang digunakan secara spesifik adalah uraian nama barang yang terdapat pada dokumen faktur pajak, PIB, dan PEB. Data ini dianalisis untuk mengidentifikasi pola-pola transaksi yang relevan untuk menemukan topik yang tepat bagi analisis rantai pasokan industri farmasi.

Topik yang tepat dapat menggambarkan alur produksi perusahaan pada industri kimia. Hasil dari identifikasi atas transaksi perusahaan menghasilkan topik atas transaksi yang relevan dengan kegiatan usaha industri farmasi, antara lain:

1) jasa;
2) promosi dan *research and development* (R&D);
3) obat;
4) bahan kimia;
5) *lab equipment*; dan
6) barang lain.

Urutan pengelompokan topik menjadi acuan dalam penentuan prioritas topik. Sebagai contoh, apabila dalam satu transaksi tercantum topik jasa dan obat, transaksi ditandai (*flagging*) sebagai topik jasa. Begitu pula jika transaksi memuat topik obat dan *lab equipment*, sistem menetapkannya sebagai topik obat. Transaksi yang tidak termasuk

dalam kelompok topik mana pun dimasukkan ke dalam kategori barang lain.

Pembentukan *Bag of Words* (BOW) dalam model NLP bertujuan membantu model LDA mengubah teks dalam faktur pajak, PIB, dan PEB menjadi representasi numerik yang dapat dianalisis oleh algoritma *machine learning*. Dalam proses pengembangan *keywords* pada *BOW* topik obat, penulis menyadur daftar obat dari laman resmi Badan Pengawasan Obat dan Makanan (BPOM) melalui tautan https://cekbpom.pom.go.id/produk-obat. Dari sumber tersebut, diperoleh 26.271 *keywords* berdasarkan merk obat.

Sementara itu, untuk kelompok topik bahan kimia, penulis menggunakan data unsur serta senyawa kimia yang bersumber dari database *Pubchem* (https://pubchem.ncbi.nlm.nih.gov/) yang dikombinasikan dengan data bahan obat dari laman cek produk BPOM. Integrasi kedua sumber tersebut menghasilkan 1.072 *keywords* bahan kimia. Adapun untuk kategori jasa, promosi dan R&D, serta *lab equipment*, penulis melakukan eksplorasi terhadap sebaran kata dalam basis data transaksi. Melalui proses tersebut, diidentifikasi sebanyak 167 kata kunci untuk jasa, 93 kata kunci untuk promosi dan R&D, serta 176 kata kunci untuk *lab equipment*. Lebih lanjut, algoritma *Word2Vec* diterapkan untuk memperluas cakupan kata kunci pada BOW, termasuk untuk mengakomodasi variasi penulisan atau kesalahan ketik (*typo*) dalam dokumen transaksi, seperti '*paracetamol*' dan '*paractamol*'.

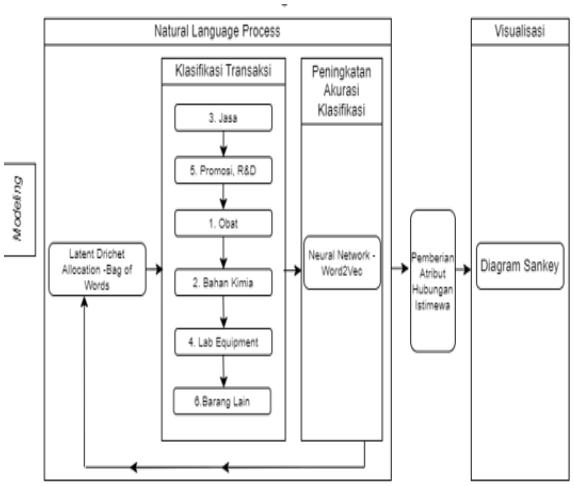
.

3.3 *Data Preparation*

Data preparation melibatkan proses pembersihan dan transformasi data untuk memastikan data siap digunakan dalam analisis. Dalam tahap pembersihan data, hal yang dilakukan adalah sebagai berikut:

1) penghapusan karakter khusus;
2) penghapusan *stop words*;
3) tokenisasi
4) *stemming* dan *lemmatization*;
5) penghapusan *whitespace* berlebih;
6) *lowercasing*;
7) penghapusan teks non-ASCII;
8) penghapusan URL dan tautan; dan
9) normalisasi teks.

Gambar 6
Alur *Modelling*



Catatan. Sumber: Diolah penulis

3.4 *Modelling*

Alur kegiatan *modelling* berjalan seperti pada gambar 6. Kegiatan *modelling* dimulai dengan menerapkan teknik NLP, seperti LDA dan algoritma *Word2Vec*. LDA digunakan untuk mengidentifikasi topik-topik dalam teks faktur pajak, PIB, dan PEB. LDA dilaksanakan menggunakan model *BOW*. BOW tersebut berisi *keywords* yang menjadi dasar bagi algoritma LDA dalam memberikan *flag* pada transaksi yang dipetakan. Pelabelan transaksi ini dilakukan hingga tingkat detail jenis obat dan bahan kimia.

Rumus algoritma LDA pada penelitian ini disadur dari penelitian Blei (2012) sebagai berikut:

$$p(\mathbf{w},\mathbf{z},\theta,\beta|\alpha,\eta)=\prod_{d=1}^{M}p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}p(z_{d,n}|\theta_d)p(w_{d,n}|z_{d,n},\beta)\right)\prod_{k=1}^{K}p(\beta_k|\eta \quad (1)$$

Di mana:

a. $p(\theta_d|\alpha)$ adalah probabilitas distribusi topik untuk dokumen d diberikan parameter *Dirichlet* α.
b. $p(z_{d,n}|\theta_d)$ adalah probabilitas topik untuk kata ke-nnn dalam dokumen ddd diberikan distribusi topik θd .

c. $p(w_{d,n}|z_{d,n},\beta)$ adalah probabilitas kata ke-nnn dalam dokumen ddd diberikan topik zd,nz_{d,n}zd,n dan distribusi kata β.

d. $p(\beta_k|\eta)$ adalah probabilitas distribusi kata untuk topik kkk diberikan parameter *Dirichlet* η.

e. *Dirichlet* adalah distribusi probabilitas multivariat.

Penulis menggunakan pendekatan *skip-gram* pada rumus algoritma *Word2Vec*. Pendekatan *skip-gram* dipilih dikarenakan untuk memaksimalkan tingkat derajat kedekatan antarpotongan kata dari transaksi yang ada. Melalui pendekatan ini, penulisn dapat melihat *keywords* baru ataupun *keywords* yang terdapat salah penulisan pada transaksi perusahaan yang belum terdapat pada *BOW*. Formula statistik *Word2Vec skip-gram* menggunakan pendekatan *negative-sampling* sebagaimana tertera pada penelitian Mikolov et al. (2013) dengan gambaran sebagai berikut.

$$\mathcal{L} = \sum_{t=1}^{T} \sum_{-m \le j \le m, j \neq 0} \left(\log \sigma(\mathbf{v}'_{w_{t+j}} \cdot \mathbf{v}_{w_t}) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} [\log \sigma(-\mathbf{v}'_{w_i} \cdot \mathbf{v}_{w_t}) \right. \quad (2)$$

Di mana:

$\sigma(x) = \frac{1}{1+\exp(-x)}$ adalah fungsi *Sigmoid*.

wt adalah kata target pada posisi *t*.

wt+j adalah kata konteks yang berada dalam konteks sekitar *wt*.

$\mathbf{v}'_{w_{t+j}}$ adalah vektor output dari kata konteks wt+j.

$\mathbf{v}_{w_t}$ adalah vektor input dari kata target wt.

wi adalah kata-kata yang diambil dari distribusi negatif Pn(w).

$\mathbf{v}'_{w_i}$ adalah vektor output dari kata negatif wi.

"log p(w)" merujuk pada logaritma dari probabilitas kata tersebut dalam konteks yang diberikan.

Setelah dilakukan pemetaan topik, data diberikan atribut relasi hubungan istimewa perusahaan dari sudut pandang perpajakan (Pasal 18 ayat (4) Undang-Undang Nomor 36 Tahun 2008 tentang Pajak Penghasilan (PPh). Pada bagian akhir, data hasil dari pemetaan dan pemberian atribut divisualisasikan pada diagram Sankey. Visualisasi digambarkan dalam dua alur, yaitu *node* serta *link*. Alur *node* dirumuskan sebagai berikut (Otto et al., 2022):

$$\mathcal{N} = \bigcup_{d \in \mathcal{D}} \{\text{source}(d), \text{target}(d)\} \quad (3)$$

dengan $\mathcal{N}$ adalah *node*. *Node* berisi perusahaan, status hubungan istimewa, dan jenis transaksi perusahaan.

Alur *link* dirumuskan sebagai berikut (Otto et al., 2022):

$$\mathcal{L} = \{(\text{source}(d), \text{target}(d), \text{Jumlah}(t)) \mid t = (d, \text{Jumlah}) \in \mathcal{T}\} \quad (4)$$

dengan $\mathcal{L}$ adalah *link*. *Link* berisi nilai transaksi antarperusahaan.

3.5 Evaluasi

Proses evaluasi bertujuan untuk mengukur kinerja dan akurasi model yang telah dibangun. Evaluasi ini mencakup dua kegiatan utama, yaitu *perplexity* dan evaluasi ekstrinsik.

Perplexity adalah metrik yang digunakan untuk mengukur seberapa baik model memprediksi data. Semakin rendah nilai *perplexity*, semakin baik model dalam memprediksi data tersebut (Belinkov & Glass, 2019). *Perplexity* digunakan untuk menilai model LDA yang mengidentifikasi topik dalam dokumen faktur pajak, PIB, dan PEB. Nilai *perplexity* menggambarkan jumlah transaksi yang belum teridentifikasi jenis topiknya. Semakin rendah nilai *perplexity*, maka semakin sedikit jumlah transaksi yang belum teridentifikasi jenis topiknya. Rumus *perplexity* untuk model LDA adalah sebagai berikut:

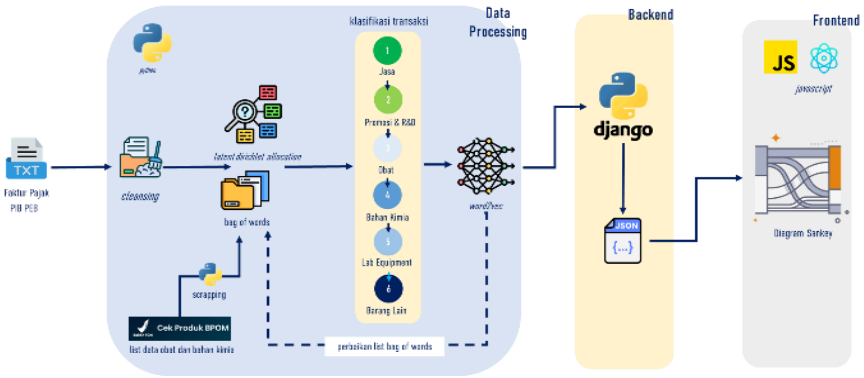
$$\text{Perplexity}(D) = \exp\left(-\frac{\sum_{d=1}^{D} \log P(w_d)}{\sum_{d=1}^{D} N_d}\right) \quad (5)$$

di mana:

- D merupakan jumlah populasi transaksi yang diuji dalam penelitian ini,
- P(wd) adalah probabilitas topik dalam populasi menurut model LDA
- Nd adalah jumlah topik dalam populasi transaksi d.

Penulis menetapkan batasan nilai *perplexity* pada angka 30 persen dari total transaksi yang

Gambar 7
Arsitektur Prototipe



Catatan. Sumber: Diolah penulis

dimodelkan (Blei, 2012). Oleh karena itu, model terus diperbaiki melalui pengayaan kata kunci (*key-words*) maupun perluasan klaster berdasarkan nilai *cosine similarity* pada *Word2Vec* hingga mampu menggambarkan topik minimal 70 persen dari total transaksi. Perbaikan tersebut dilakukan salah satunya dengan memperkaya kata kunci pada *BOW*.

Evaluasi ekstrinsik dilakukan untuk menilai kinerja model dalam aplikasi dunia nyata. Evaluasi ekstrinsik dilakukan dengan dua pendekatan, yaitu evaluasi kinerja analisis rantai pasok berdasarkan sisi *realibility* dan sisi *cost*.

Dari sisi analisis *reliability* rantai pasok, penulis menilai apakah visualisasi Sankey dan topik yang telah dipetakan mampu menunjukkan kesesuaian antara barang *output* dan *input* perusahaan. Sementara itu, dari sisi analisis *cost* rantai pasok, penulis meninjau kemampuan visualisasi tersebut dalam membantu pengguna mengidentifikasi pengeluaran yang tidak sesuai dengan proses bisnis perusahaan di industri farmasi.

Berdasarkan penelitian Qiu et al. (2018), pelaksanaan evaluasi intrinsik berkorelasi dengan hasil evaluasi ekstrinsik sehingga kedua proses tersebut tidak dapat dipisahkan. Oleh karena itu, evaluasi intrinsik melalui nilai *perplexity* dan evaluasi ekstrinsik dilaksanakan secara beriringan. Apabila penulis menganggap model telah memenuhi kriteria evaluasi intrinsik dan ekstrinsik, penelitian dilanjutkan ke tahap implementasi (*deployment*) prototipe.

3.6 Evaluasi

Tahapan terakhir dalam metodologi CRISP-DM adalah *deployment*. Tahap *deployment* dilakukan dalam bentuk prototipe yang ditayangkan melalui aplikasi web melalui *React Java Script* seperti pada gambar 7.

Secara alur, prototipe ini dimulai dengan penyiapan data menggunakan bahasa pemrograman Python. Penulis membersihkan data mentah terlebih dahulu, kemudian menentukan topik transaksi menggunakan model NLP LDA-BOW yang disempurnakan dengan *Word2Vec*. Hasil pemetaan tersebut dibaca oleh *backend* aplikasi web menggunakan Django. Pada bagian *backend*, dataset disalurkan dalam format JSON yang terstruktur menjadi satuan *node* dan *link* untuk dibaca oleh visualisasi diagram Sankey pada *frontend*. Penelitian ini menggunakan JavaScript untuk menampilkan visualisasi tersebut dengan memanfaatkan paket *Apache ECharts* dari *Apache Software Foundation*.

4. HASIL DAN PEMBAHASAN

4.1 Ekstraksi Topik Transaksi

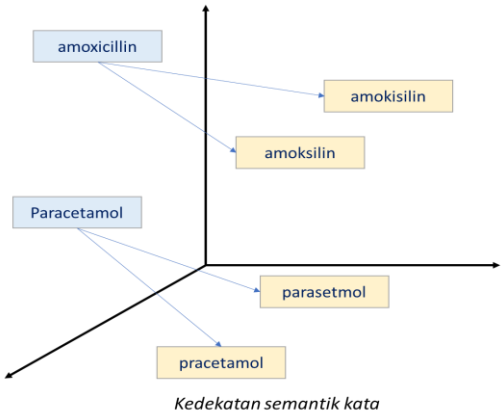
Penerapan model NLP LDA-BOW pada 6.897.594 dokumen transaksi Wajib Pajak (dokumen pembelian dan penjualan faktur pajak, PIB-PEB) pada bidang industri farmasi, menghasilkan sebaran topik seperti pada tabel 1. Sementara itu,

Tabel 1
Sebaran Topik Awal

No	Jenis Topik	Jumlah Merek	Persentase Transaksi
1	Obat	5.587	52,70%
2	Bahan Kimia	225	3,10%
3	Jasa	66	4,40%
4	Marketing & Research	49	1,24%
5	Lab Equipment	57	0,09%
6	Barang Lainnya (Tidak teridentifikasi)	–	39,40%

Catatan. Sumber: Diolah penulis

Gambar 8
Ilustrasi Kedekatan Semantik Kata



Catatan. Sumber: Diolah penulis

penerapan algoritma *Word2Vec* pada LDA-BOW, menghasilkan sebaran topik pada tabel 2.

Pada Tabel 2, terdapat peningkatan jumlah transaksi yang teridentifikasi topiknya sebesar 19,07%, yakni dari 4.244.090 transaksi pada Tabel 1 menjadi 5.053.828 transaksi. Jumlah merek yang teridentifikasi juga mengalami kenaikan sebesar 18,5%, dari 5.984 jenis merek menjadi 7.089 merek. Selain kenaikan tersebut, penulis mendefinisikan 10 kelompok topik baru (baris berwarna biru muda) yang dibentuk melalui model *clustering* dengan memanfaatkan nilai *cosine similarity* antarvektor kata pada dokumen transaksi. Kata-kata yang memiliki nilai *similarity* tinggi dikelompokkan ke dalam klaster yang sama. Hal ini sejalan dengan penelitian Sandhu et al. (2023) bahwa teknik

Tabel 2
Sebaran Topik Setelah Integrasi Word2Vec

No	Jenis Topik	Jumlah Merek	Persentase Transaksi
1	Obat	6.337	58,33%
2	Bahan Kimia	315	5,81%
3	Jasa	106	4,60%
4	Marketing & Research	57	1,67%
5	Lab Equipment	67	0,11%
6	Barang Lain Konstruksi	17	1,4226%
7	Barang Lain Tanah	5	0,3962%
8	Barang Lain Mesin	18	0,3385%
9	Barang Lain Bahan Bakar	7	0,1870%
10	Barang Lain ATK	65	0,1433%
11	Barang Lain Kendaraan	70	0,1156%
12	Barang Lain Intangible	5	0,0711%
13	Barang Lain Travel	8	0,0332%
14	Barang Lain Handphone	4	0,0182%
15	Barang Lain Perhiasan	8	0,0163%
16	Barang Lainnya (Tidak teridentifikasi)	–	26,70%

Catatan. Sumber: Diolah penulis

clustering berbasis *cosine similarity* dapat dimanfaatkan untuk mengekstraksi topik dari sebuah teks. Pemanfaatan *Word2Vec* dalam perluasan topik ini berhasil menangkap merek baru meskipun terdapat kesalahan penulisan (*typo*) dalam dokumen transaksi, sebagaimana tertera pada Tabel 3 dan diilustrasikan pada Gambar 8.

Hasil perluasan merek dari *Word2Vec* dimanfaatkan untuk memperkaya *Bag of Words* pada LDA-BOW guna memperluas cakupan ekstraksi topik. Adapun sebaran topik untuk

seluruh transaksi perusahaan disajikan pada Tabel 4, 5, 6,dan 7.

Tabel 3
Contoh Merek Salah Tulis Hasil Perluasan Merek Word2Vec

Topik	Merek	Perluasan Merek dari Word2Vec
Obat	paracetamol	parasetmol
		paractamol
		pracetamol
	caffeine	cafeine
		cafein
	amoxicillin	amoxicilin
		amoksilin
		amokisilin
	ibuprofen	ibuprofn
		ibupofen

Catatan. Sumber: Diolah penulis

Tabel 4
Sebaran Jenis Topik Obat*

No	Jenis Merek	Jumlah	%
1	Vitamin & Suplemen A	95.857	2,38%
2	Obat Demam A	34.003	0,85%
3	Obat Demam B	24.862	0,62%
4	Minyak Angin A	22.490	0,56%
5	Obat Tradisional A	22.293	0,55%
6	Minyak Angin B	21.355	0,53%
7	Cairan Infus A	19.435	0,48%
8	Obat Asam Lambung	15.485	0,38%
9	Obat Darah Tinggi A	13.902	0,35%
10	Obat Kosmetik A	13.777	0,34%
11	Minyak Angin C	13.191	0,33%
12	Merek Obat Lain	3.727.031	92,63%

*jenis merek obat telah dimasking
Catatan. Sumber: Diolah penulis

Tabel 5
Sebaran Jenis Topik Jasa

No	Jenis Merek	Jumlah	%
1	Jasa	102.041	32,2%
2	Fee	62.581	19,7%
3	Biaya	27.758	8,8%
4	Charge	20.007	6,3%
5	Sewa	19.140	6,0%
6	Charges	13.557	4,3%
7	Handling	11.543	3,6%
8	Service	10.743	3,4%
9	Pengiriman	10.405	3,3%
10	Brosur	10.093	3,2%
11	Kalibrasi	6.157	1,9%
12	Merek Jasa Lain	23.023	7,3%

Catatan. Sumber: Diolah penulis

Tabel 6
Sebaran Jenis Topik Bahan Kimia

No	Jenis Merek	Jumlah	%
1	Sodium	16.995	4,24%
2	Liquid	16.420	4,10%
3	Methyl	16.408	4,09%
4	Monohydrate	14.202	3,54%
5	Alkohol	9.521	2,38%
6	Hydrate	6.898	1,72%
7	Chloride	6.336	1,58%
8	Menthol	5.110	1,27%
9	Sulfat	4.586	1,14%
10	Sucralfat	4.500	1,12%
11	Potassiu	3.967	0,99%
12	Merek Bahan Kimia Lain	295.936	73,82%

Catatan. Sumber: Diolah penulis

Tabel 7
Sebaran Jenis Topik Barang Lain Teridentifikasi*

No	Jenis Merek	Jumlah	%
1	Hotel/Penginapan	17.502	9,25%
2	Storage	10.930	5,78%
3	Document	6.385	3,38%
4	Kertas	4.944	2,61%
5	Airport Fee/Tiket	4.838	2,56%
6	Tinta	3.102	1,64%
7	Amplop	2.276	1,20%
8	Spidol	2.181	1,15%
9	Binder	2.169	1,15%
10	Printer	2.168	1,15%
11	Buku	1.660	0,88%
12	Merek Barang Lain Lainnya	130.974	69,25%

*barang lain yang terdapat pada topik (konstruksi, tanah, mesin, bahan bakar, atk, kendaraan, intangible, travel, handphone, & perhiasan)
Catatan. Sumber: Diolah oleh penulis.

4.2 Analisis Rantai Pasokan dan Visualisasi Transaksi

Analisis rantai pasokan difokuskan pada dua aspek yaitu keandalan (*reliability*) dan biaya (*cost*). Untuk menjawab tantangan tersebut, penulis telah menyiapkan empat jenis visualisasi dalam prototipe aplikasi guna melakukan analisis tersebut. Keempat jenis visualisasi tersebut adalah:

a. Visualisasi Overview Transaksi Perusahaan;
b. Visualisasi *Supplier-Customer*;
c. Visualisasi Alur Produksi Merek; dan
d. Visualisasi Detil Topik.

Visualisasi pertama merupakan overview transaksi perusahaan yang disajikan pada Lampiran A, yaitu Visualisasi Overview Transaksi Perusahaan Indonesia Oke Jaya. Pada visualisasi ini, penulis menggunakan diagram Sankey untuk menggambarkan transaksi perusahaan secara umum dengan memadukan jenis topik pembelian-penjualan. Alur tersebut kemudian disandingkan dengan informasi mengenai ada atau tidaknya hubungan istimewa dalam transaksi tersebut. Pada *node* terakhir, penulis menampilkan lokasi sumber pembelian maupun tujuan penjualan, baik di dalam negeri maupun luar negeri. Melalui visualisasi ini, pengguna dapat mengidentifikasi asal dan tujuan transaksi, jenis komoditas yang ditransaksikan, serta relasi hubungan istimewa antara perusahaan dengan pelanggan (*customer*) maupun pemasok (*supplier*).

Visualisasi kedua yang tertera pada Lampiran B menyajikan Visualisasi Alur *Supplier-Customer* Perusahaan Indonesia Oke Jaya. Pada tahap ini, penulis menampilkan detail pemasok dan pelanggan yang disandingkan dengan jenis topik transaksinya. Penulis membedakan perusahaan lawan transaksi yang memiliki hubungan istimewa dengan pemberian warna kuning pada *node*, sedangkan perusahaan tanpa hubungan istimewa menggunakan warna biru. Selain itu, *node* tersebut juga menampilkan bidang usaha dari masing-masing pihak terkait. Untuk menjaga kejelasan visual, penulis hanya menampilkan peringkat 20 besar berdasarkan nilai transaksi, sementara perusahaan di bawah peringkat tersebut digabungkan ke dalam satu *node* kolektif.

Selanjutnya, pada Lampiran C, penulis menyertakan fitur *hover* yang aktif saat salah satu *node* dipilih. Sebagai contoh, ketika *node* pembelian jasa dipilih, fitur ini menampilkan aliran pembelian jasa yang berasal dari berbagai pemasok secara spesifik. Melalui ketiga visualisasi awal ini, pengguna diharapkan dapat memahami alur transaksi secara menyeluruh serta mampu menakar area risiko yang memerlukan penelitian atau pemeriksaan lebih lanjut.

Pada Lampiran D (Visualisasi Alur Produksi Obat Demam XYZ), penulis memvisualisasikan alur produksi dengan memanfaatkan data komposisi obat dari laman BPOM. Melalui visualisasi ini, pengguna dapat mengamati bahan baku atau bahan kimia penyusun obat tersebut beserta besaran nilainya. Lebih lanjut, pengguna dapat meninjau rekanan *supplier* maupun *customer* perusahaan dalam pengadaan bahan baku, termasuk mengidentifikasi ada tidaknya hubungan istimewa. Dalam visualisasi ini, terlihat bahwa perusahaan menjual seluruh hasil produksi obat

demam XYZ kepada distributor yang merupakan pihak afiliasi. Perlu dicatat bahwa data bahan kimia dalam proses pembuatan obat ini telah dilakukan penyamaran (*masking*) oleh penulis demi menjaga kerahasiaan data.

Selanjutnya, pada Lampiran E (Visualisasi Alur Produksi Obat Demam XYZ Detail Pembelian Paracetamol), penulis menyertakan fitur *hover*. Pengguna dapat mengeklik salah satu komponen bahan baku (misalnya paracetamol) untuk memunculkan aliran detail pembelian bahan tersebut secara spesifik.

Pada bagian akhir, penulis menyajikan visualisasi detail topik jasa pada Lampiran F. Melalui gambar tersebut, penulis menggambarkan detail merek dalam suatu topik dengan menampilkan perusahaan lawan transaksi (*pembelian-penjualan*). Sebagai contoh, pada topik pembelian jasa, pengguna dapat melihat jenis jasa yang dibeli, status hubungan istimewa lawan transaksi, hingga bidang usaha penyedia jasa guna mendapatkan gambaran yang komprehensif. Fungsi *hover* pada topik ini juga diperlihatkan pada Lampiran G, yang memungkinkan pengguna melihat jenis transaksi jasa apa saja yang dilakukan oleh *supplier* dengan hubungan istimewa terhadap perusahaan terkait.

Terkait dengan analisis rantai pasok, penulis berpendapat bahwa penggunaan keempat jenis visualisasi tersebut dapat mendukung analisis dari perspektif keandalan (*reliability*) dan biaya (*cost*). Dari sisi keandalan, visualisasi ini membantu pengguna (khususnya fiskus) mengidentifikasi titik-titik topik pembelian dan penjualan untuk memastikan relevansi transaksi dengan kegiatan usaha. Pengguna dapat menghitung persentase jumlah topik "barang lain" terhadap total pembelian dan membandingkannya dengan rata-rata industri. Persentase "barang lain" yang besar mengindikasikan risiko tinggi pada keandalan rantai pasok. Jika ditemukan ketidaksesuaian antara pembelian bahan baku dengan barang jadi yang dijual, hal tersebut dapat dijadikan area penelitian atau pemeriksaan lebih lanjut untuk mendeteksi adanya transaksi yang tidak dilaporkan dalam faktur pajak maupun PIB/PEB.

Analisis rantai pasok dari sisi biaya (*cost*) memungkinkan pengguna mengidentifikasi titik pengeluaran perusahaan yang paling signifikan, yang dapat dimulai dari peninjauan topik jasa serta barang lain. Melalui Lampiran D, pengguna dapat menganalisis apakah volume obat yang diproduksi telah sebanding dengan jumlah bahan baku yang dibeli berdasarkan detail merek. Lebih lanjut, dengan memanfaatkan data hubungan istimewa, pengguna dapat membandingkan besaran nilai transaksi jasa atau barang antara pihak afiliasi dengan pihak independen. Titik ini menjadi dasar bagi pengguna untuk memperdalam penelitian atau pemeriksaan guna menguji kepatuhan perpajakan perusahaan.

Pemanfaatan informasi hubungan istimewa sangat krusial dalam pengujian kepatuhan perpajakan. Hal ini sejalan dengan penelitian Park (2018) yang menyatakan bahwa perusahaan dalam kelompok bisnis cenderung menggunakan transaksi pihak terkait sebagai modus penghindaran pajak. Berdasarkan implementasi keempat jenis visualisasi tersebut, penulis berpendapat bahwa analisis rantai pasok dari perspektif keandalan (*reliability*) serta biaya (*cost*) sangat dimungkinkan untuk dilakukan dalam ruang lingkup perpajakan.

Prototipe dalam penelitian ini memiliki potensi untuk diduplikasi pada sektor audit internal maupun eksternal guna menentukan area kritis audit. Selain itu, penelitian ini diharapkan dapat menjadi *blueprint* metodologis bagi studi-studi lanjutan mengenai deteksi *trade misinvoicing* dengan menggunakan teknik NLP.

5. KESIMPULAN

Penelitian ini menunjukkan bahwa penerapan LDA BOW serta *Word2Vec* efektif dalam mengidentifikasi dan mengklasifikasikan topik transaksi faktur pajak serta PIB-PEB yang memiliki penulisan tidak baku. Metode ini berhasil memetakan 73,7% transaksi. Hal ini sejalan dengan hasil penelitian yang dilakukan oleh Li et al. (2018) bahwa penggabungan model LDA dengan *Word2Vec* secara efektif dapat melakukan ekstraksi kalimat pada teks pendek serta domain yang sempit. Lebih lanjut penelitian yang dilakukan oleh Li et al. (2018) juga menyebutkan bahwa *Word2Vec* secara efektif dapat meningkatkan akurasi kinerja

ekstraksi topik atas hasil dari LDA. Senada dengan hasil penelitian ini, Saout et al. (2024) menyatakan bahwa penerapan model NLP pada dimanfaatkan dalam deteksi jenis transaksi pada faktur.

Dari sisi visualisasi, pemanfaatan Diagram Sankey dalam analisis rantai pasokan dapat dengan baik menggambarkan aliran transaksi perusahaan, hal ini sejalan dengan salah satu hasil penelitian Rudolf & Martina (2019) bahwa pemanfaatan diagram Sankey menggambarkan alur aliran material produksi serta titik biaya perusahaan untuk dimanfaatkan dalam analisis biaya perusahaan. Akhir kata, pemanfaatan teknik NLP dan visualisasi Sankey dapat membantu fiskus dan pengguna untuk melakukan analisis rantai pasokan di bidang perpajakan mempelajari alur barang dan jasa, sehingga memungkinkan deteksi yang lebih cepat dan akurat terhadap area risiko perusahaan. Secara keseluruhan, penggunaan model NLP dan visualisasi Sankey dalam analisis rantai pasokan dapat memberikan alat yang efektif dan efisien bagi fiskus untuk mengelola dan mengoptimalkan kegiatan pengawasan atas rantai pasokan Wajib Pajak dari perspektif keandalan maupun biaya.

6. IMPLIKASI DAN KETERBATASAN

Integrasi model *Natural Language Processing* (NLP) seperti *Latent Dirichlet Allocation Bag of Words* (LDA BOW) dan *Word2Vec* dengan visualisasi diagram Sankey pada analisis rantai pasokan menawarkan cara yang lebih efisien dan efektif untuk mengidentifikasi risiko dan potensi penghindaran pajak dalam pengujian rantai pasokan perusahaan. Hal ini sejalan dengan tujuan modernisasi administrasi perpajakan yang lebih responsif dan adaptif terhadap tantangan digitalisasi (OECD, 2021). Penulis berharap hasil dan teknik penelitian melalui metode CRISP-DM dapat direplikasi di dunia audit keuangan serupa yang membutuhkan kegiatan analisis aliran barang jasa dari dokumen transaksi. Keterbatasan dari penelitian ini adalah generalisasi hasil penelitian ini ke industri lain, karakteristik dan kompleksitas data yang berbeda antara satu industri dengan yang lain, memerlukan penyiapan data serta pembangunan *keywords* pada *Bag of Words* yang berbeda. Penelitian lebih lanjut dapat meneliti pengaruh variasi *keywords* dan parameter model (jumlah topik, ukuran konteks Word2Vec) terhadap *perplexity* dan kinerja klasifikasi dalam konteks klasifikasi jenis barang dan jasa. Di lain sisi, penulis juga berharap penelitian lebih lanjut dapat mengeksplorasi model *large language model* (LLM) dalam *topic modelling* untuk meningkatkan cakupan dan ketepatan klasifikasi barang dan jasa.

DAFTAR REFERENSI

Alarie, B., Niblett, A., & Yoon, A. H. (2018). How artificial intelligence will affect the practice of law. *University of Toronto Law Journal*, *68*(1), 106–124. https://doi.org/10.3138/utlj.2017-0052

Apache Software Foundation. (n.d.). *Sankey diagram*. Apache ECharts. Diakses pada 30 Juli 2024 dari https://echarts.apache.org/examples/en/index.html#chart-type-Sankey

Badan Pengawasan Obat dan Makanan. (n.d.). *Daftar produk obat* [Data set]. https://cekbpom.pom.go.id/produk-obat

Beamon, B. M. (1998). Supply chain design and analysis: Models and methods. *International Journal of Production Economics*, *55*(3), 281–294. https://doi.org/10.1016/S0925-5273(98)00079-6

Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, *7*, 49–72. https://doi.org/10.1162/tacl_a_00254

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Cen, L., Maydew, E. L., Zhang, L., & Zuo, L. (2016). Customer–supplier relationships and corporate tax avoidance. *Journal of Accounting Research*, *54*(1), 75–114. https://doi.org/10.1016/j.jfineco.2016.09.009

Chakkarwar, V., & Tamane, S. C. (2020). Quick insight of research literature using topic modeling. In Y. D. Zhang, J. Mandal, C. So-In, & N. Thakur (Eds.), *Smart trends in computing and communications: Smart innovation, systems and technologies* (Vol. 165, pp. 189–197). Springer. https://doi.org/10.1007/978-981-15-0077-0_20

Direktorat Jenderal Pajak. (2012). *Peraturan Direktur Jenderal Pajak Nomor PER-24/PJ/2012 tentang Bentuk, Ukuran, Tata Cara Pengisian Keterangan, dan Penyampaian Surat Pemberitahuan Masa Pajak Pertambahan Nilai dan Pajak Penjualan atas Barang Mewah*.

Direktorat Jenderal Pajak. (2014). *Peraturan Direktur Jenderal Pajak Nomor PER-16/PJ/2014 tentang Tata Cara Pembuatan dan Pelaporan Faktur Pajak Berbentuk Elektronik*.

Direktorat Jenderal Pajak. (2023). *Peraturan Direktur Jenderal Pajak Nomor PER-11/PJ/2023 tentang Perubahan Kedua atas Peraturan Direktur Jenderal Pajak Nomor PER-03/PJ/2022 tentang Faktur Pajak*.

Direktorat Jenderal Pajak. (2024). *Peraturan Direktur Jenderal Pajak Nomor PER-03/PJ/2024 tentang Perubahan Ketiga atas Peraturan Direktur Jenderal Pajak Nomor PER-03/PJ/2022 tentang Faktur Pajak*.

Hugos, M. H. (2018). *Essentials of supply chain management* (4th ed.). John Wiley & Sons.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall.

Li, C., Lu, Y., Wu, J., Zhang, Y., Xia, Z., Wang, T., Yu, D., Chen, X., Liu, P., & Guo, J. (2018). LDA meets Word2Vec: A novel model for academic abstract clustering. *Proceedings of the 2018 Web Conference Companion (WWW '18)*, 369–373. https://doi.org/10.1145/3184558.3191629

Lupton, R. C., & Allwood, J. M. (2017). Hybrid Sankey diagrams: Visual analysis of multidimensional data for understanding resource use. *Resources, Conservation and Recycling*, *124*, 141–151. https://doi.org/10.1016/j.resconrec.2017.05.002

Madala, V. (2023). Topic modeling and text clustering using LDA and Word2Vec. *Journal of Machine Learning Research*, *24*(1), 44–59.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv. https://arxiv.org/abs/1301.3781

National Library of Medicine. (n.d.). *PubChem database*. Diakses pada 24 Mei 2024 dari https://pubchem.ncbi.nlm.nih.gov/

Organisation for Economic Co-operation and Development. (2004). *Compliance risk management: Managing and improving tax compliance*. OECD Publishing.

Organisation for Economic Co-operation and Development. (2021). *Tax administration 2021: Comparative information on OECD and other advanced and emerging economies*. OECD Publishing. https://doi.org/10.1787/7e3d52b0-en

Otto, E., Culakova, E., Meng, S., Zhang, Z., Xu, H., Mohile, S. G., & Flannery, M. A. (2022). Overview of Sankey flow diagrams: Focusing on symptom trajectories in older adults with advanced cancer. *Journal of Geriatric Oncology*, *13*(5), 742–746. https://doi.org/10.1016/j.jgo.2021.12.017

Park, S. (2018). Related party transactions and tax avoidance of business groups. *Sustainability*, *10*(10), 3571. https://doi.org/10.3390/su10103571

Peraturan Menteri Keuangan Nomor 155/PMK.04/2022 tentang Ketentuan Kepabeanan di Bidang Ekspor. https://jdih.kemenkeu.go.id/dok/155-pmk-04-2022

Peraturan Menteri Keuangan Nomor 190/PMK.04/2022 tentang Pengeluaran Barang Impor untuk Dipakai. https://jdih.kemenkeu.go.id/dok/190-pmk-04-2022

Qiu, Y., Li, H., Li, S., Jiang, Y., Hu, R., & Yang, L. (2018). Revisiting correlations between intrinsic and extrinsic evaluations of word embeddings. In J. Liu & Q. Liu (Eds.), *Chinese computational linguistics and natural language processing based on naturally annotated big data (CCL 2018), Lecture Notes in Computer Science* (Vol. 11221, pp. 347–358). Springer. https://doi.org/10.1007/978-3-030-01716-3_29

Rostamizadeh, A., Talwalkar, A., & Wainwright, M. J. (2018). *Foundations of machine learning*. The MIT Press.

Rudolf, K., & Martina, H. (2019). Modelling a production process using a Sankey diagram and Computerized Relative Allocation of Facilities Technique (CRAFT). *Open Engineering*, *9*, 444–449. https://doi.org/10.1515/eng-2019-0043

Sandhu, A., Edara, A., Wajid, F., & Agrawala, A. (2023). *Temporal analysis on topics using Word2Vec*. arXiv. https://arxiv.org/abs/2209.11717

Saout, T., Lardeux, F., & Saubion, F. (2024). An overview of data extraction from invoices. *IEEE Access*, *12*, 19872–19886. https://doi.org/10.1109/ACCESS.2024.3360528

Supply Chain Council. (2012). *Supply Chain Operations Reference (SCOR) model* (Version 11.0).

Sürie, C., & Wagner, M. (2005). Supply chain analysis. In H. Stadtler & C. Kilger (Eds.), *Supply chain management and advanced planning* (pp. 37–63). Springer. https://doi.org/10.1007/3-540-24814-5_3

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188. https://doi.org/10.1613/jair.2934

Undang-Undang Nomor 7 Tahun 1983 tentang Pajak Penghasilan sebagaimana telah beberapa kali diubah terakhir dengan Undang-Undang Nomor 7 Tahun 2021 tentang Harmonisasi Peraturan Perpajakan.

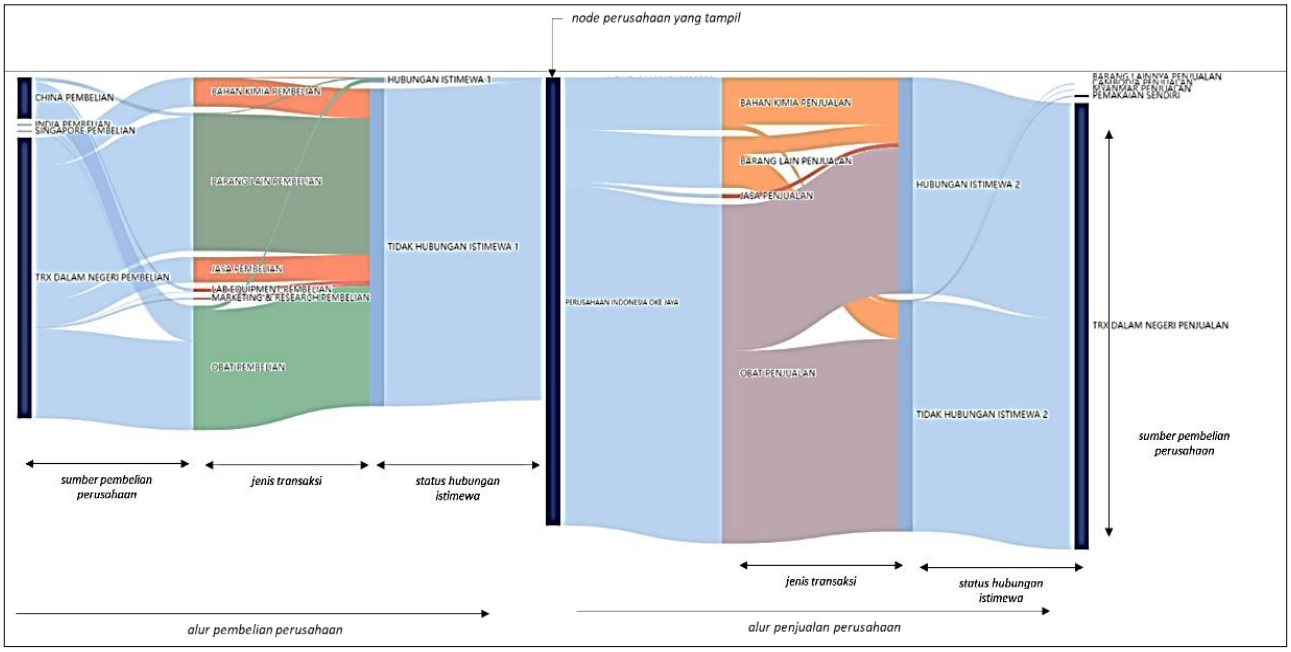
Undang-Undang Nomor 8 Tahun 1983 tentang Pajak Pertambahan Nilai Barang dan Jasa dan Pajak Penjualan atas Barang Mewah sebagaimana telah beberapa kali diubah terakhir dengan Undang-Undang Nomor 7 Tahun 2021 tentang Harmonisasi Peraturan Perpajakan.

Undang-Undang Nomor 10 Tahun 1995 tentang Kepabeanan sebagaimana telah diubah dengan Undang-Undang Nomor 17 Tahun 2006.

Undang-Undang Nomor 36 Tahun 2008 tentang Perubahan Keempat atas Undang-Undang Nomor 7 Tahun 1983 tentang Pajak Penghasilan.

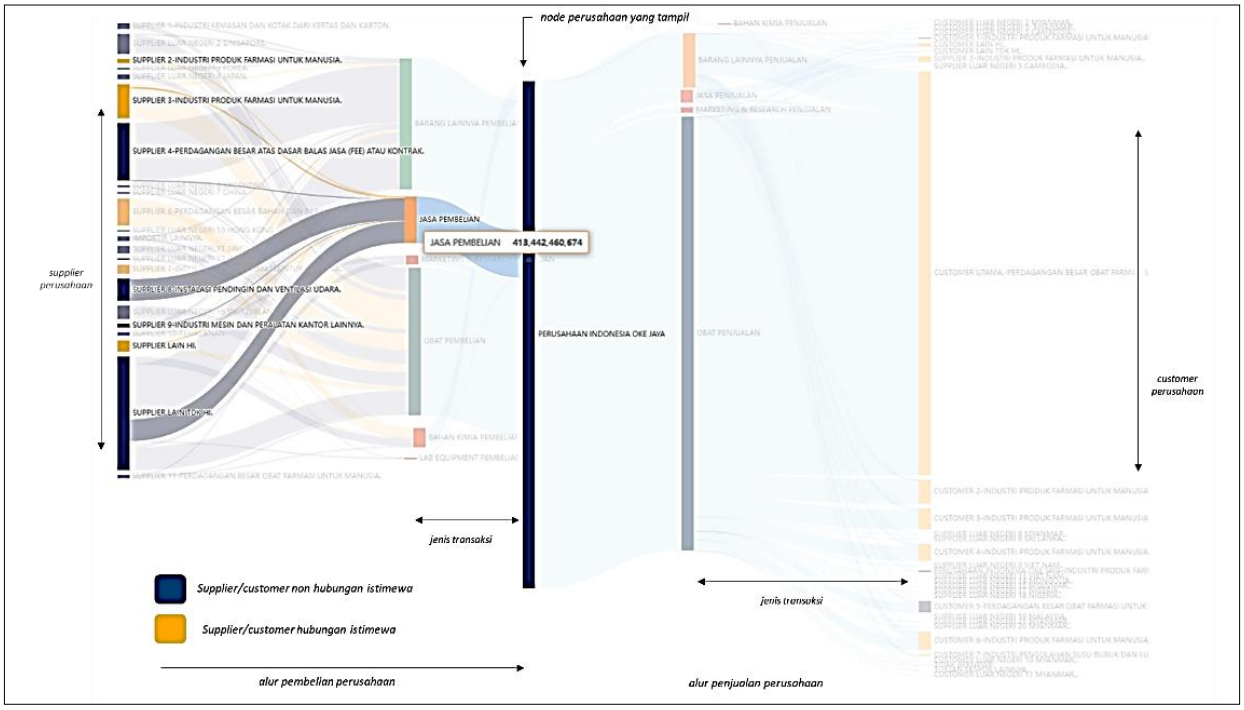
LAMPIRAN

Lampiran A
Visualisasi Overview Transaksi Perusahaan Indonesia Oke Jaya



Catatan. Sumber: diolah penulis

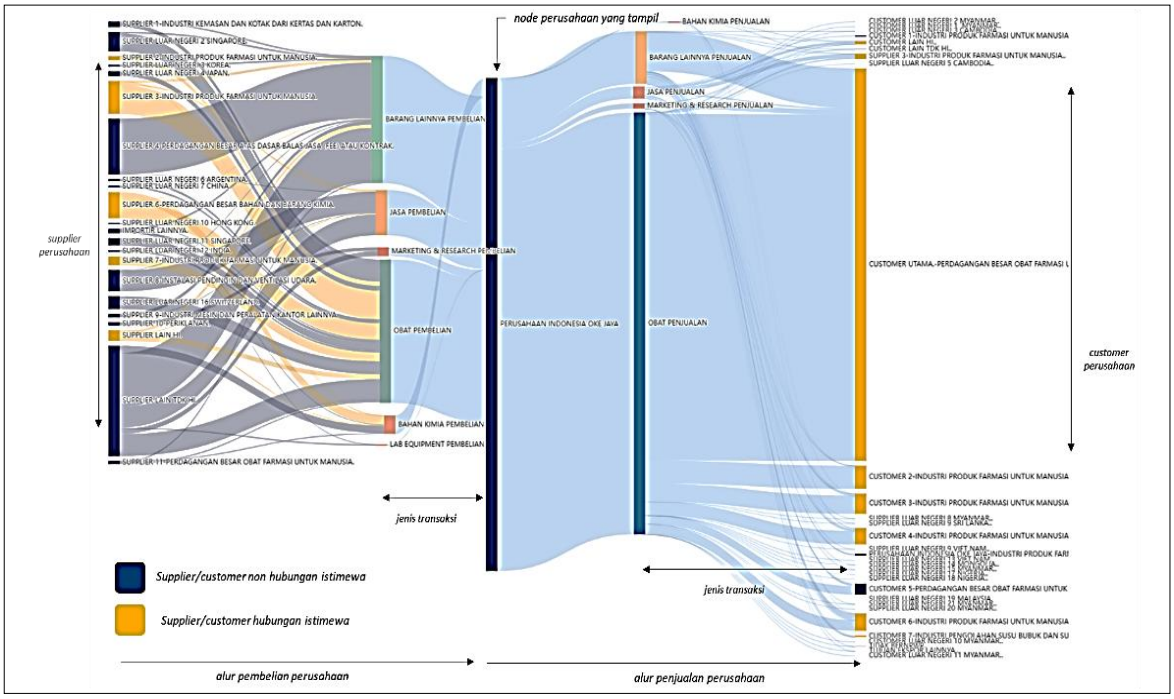
Lampiran B
Visualisasi Alur *Supplier-Customer* Perusahaan Indonesia Oke Jaya



Catatan. Sumber: Diolah Penulis

Lampiran C

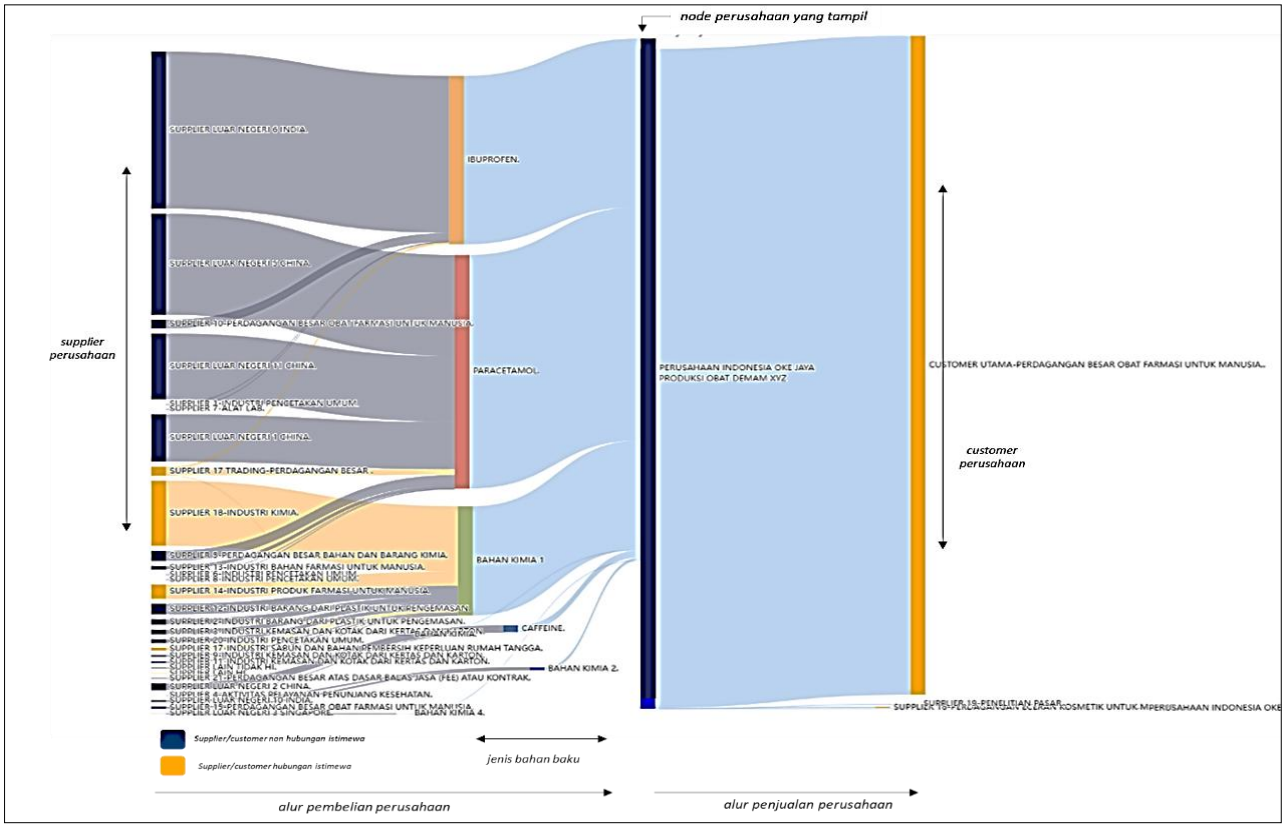
Visualisasi *Supplier-Customer* Perusahaan Indonesia Oke Jaya Klik Alur Pembelian Jasa



Catatan. Sumber: Diolah Penulis

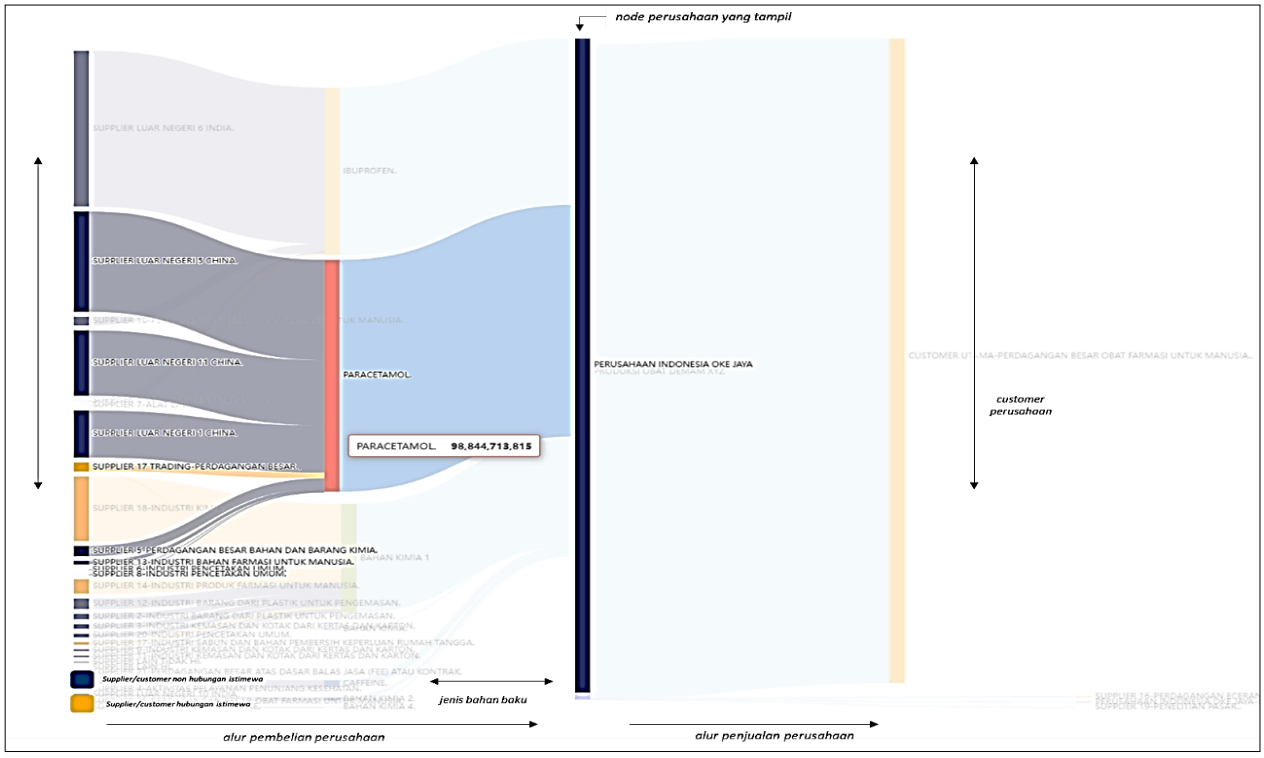
Lampiran D

Visualisasi Alur Produksi Obat Demam XYZ



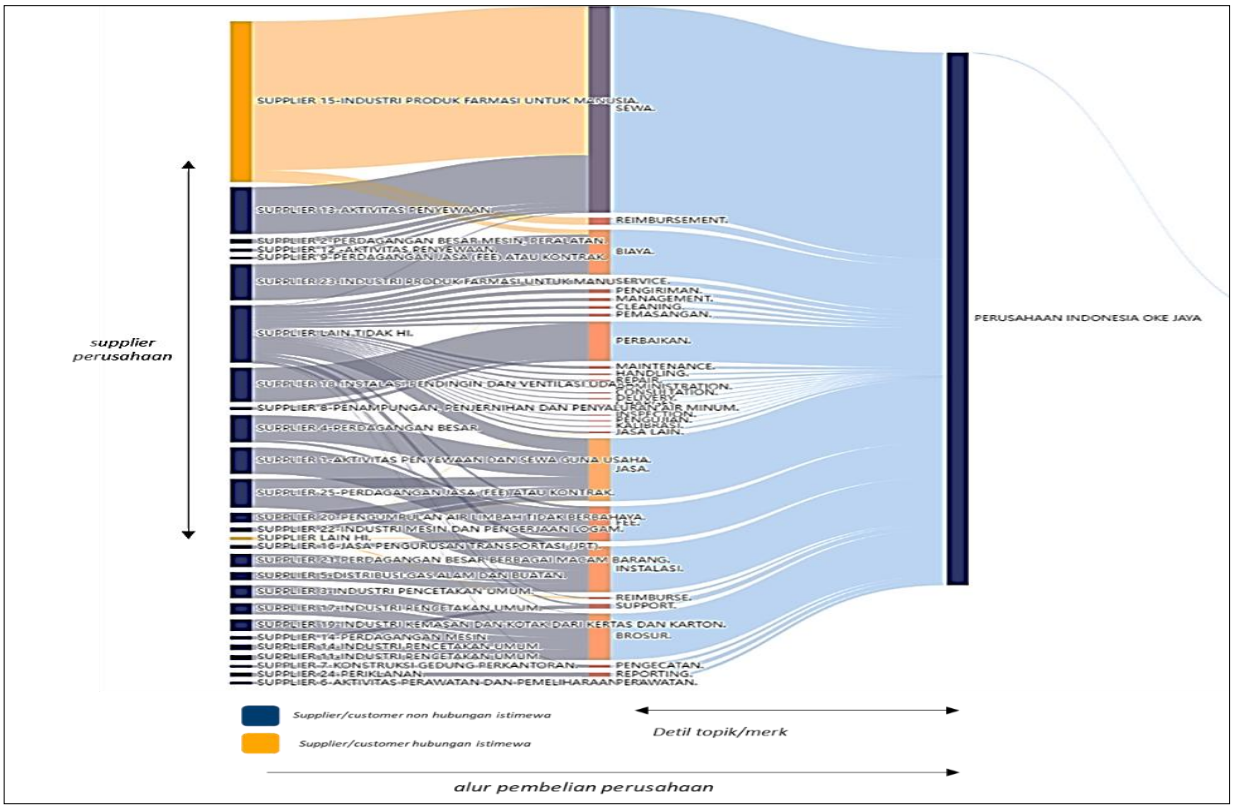
Catatan. Sumber: Diolah Penulis

Lampiran E
Visualisasi Alur Produksi Obat Demam XYZ Detail Pembelian Paracetamol



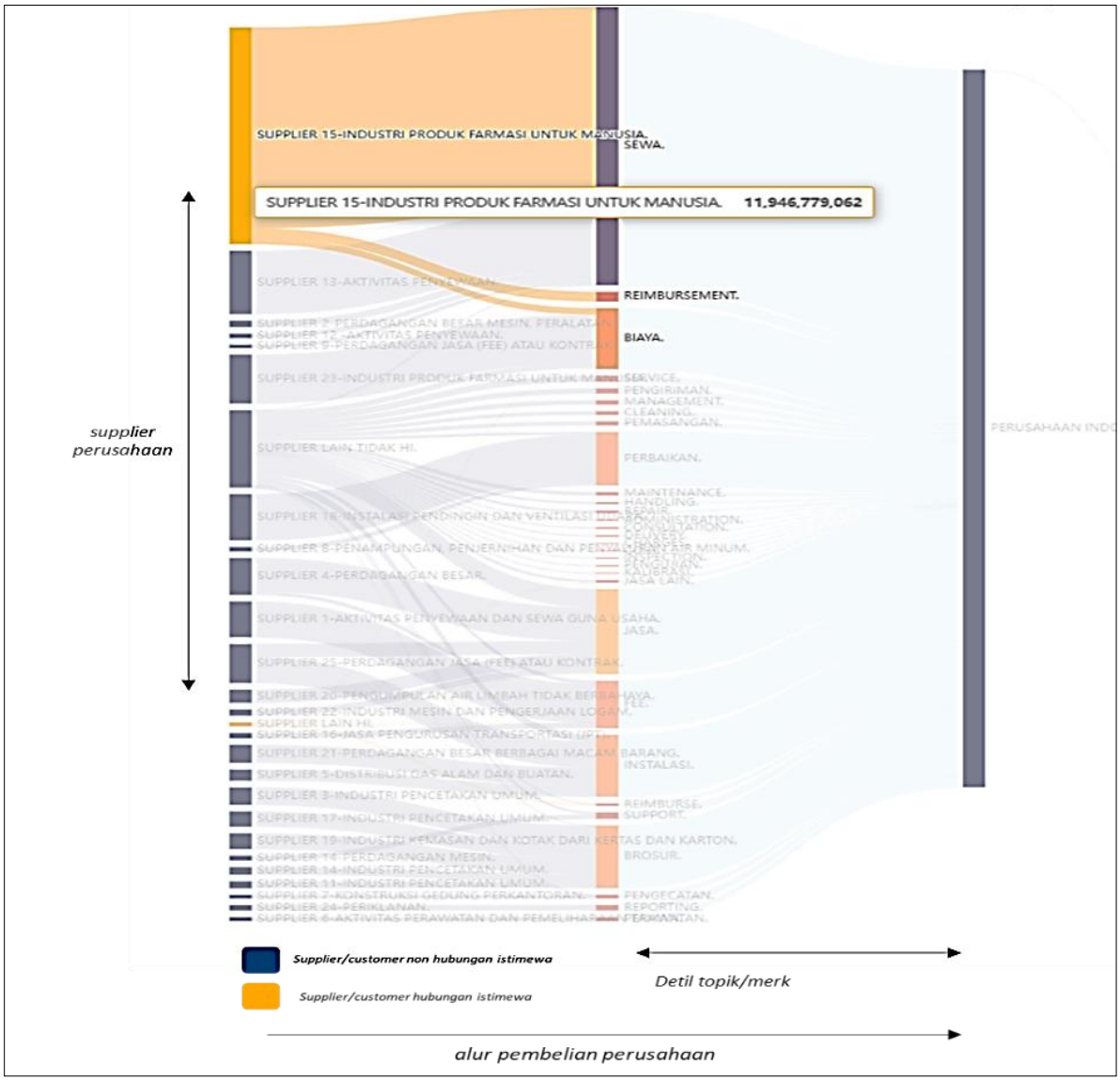
Catatan. Sumber: Diolah Penulis

Lampiran F
Visualisasi Detil Topik Jasa



Catatan. Sumber: Diolah Penulis

Lampiran G
Visualisasi Detil Topik Jasa Klik Alur *Supplier* 15-Industri Produk Farmasi



Catatan. Sumber: Diolah Penulis