ANALISIS SENTIMEN TERHADAP RENCANA KENAIKAN PAJAK PERTAMBAHAN NILAI MENJADI DUABELAS PERSEN PADA MEDIA SOSIAL X DENGAN ALGORITMA SUPPORT VECTOR MACHINE (SVM)

PROPOSAL SKRIPSI

M. ARKAN PUTRA HIMAWAN
2003040141

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNIK DAN SAINS
UNIVERSITAS MUHAMMADIYAH PURWOKERTO
2024/2025

ANALISIS SENTIMEN TERHADAP RENCANA KENAIKAN PAJAK PERTAMBAHAN NILAI MENJADI DUABELAS PERSEN PADA MEDIA SOSIAL X DENGAN ALGORITMA SUPPORT VECTOR MACHINE (SVM)

PROPOSAL SKRIPSI

diajukan sebagai syarat untuk melaksanakan penelitian dalam Mata Kuliah Skripsi

M. ARKAN PUTRA HIMAWAN
2003040141

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNIK DAN SAINS
UNIVERSITAS MUHAMMADIYAH PURWOKERTO
2024/2025

HALAMAN PERSETUJUAN

Proposal Skripsi yang diajukan oleh:

Nama	: M. Arkan Putra Himawan
NIM	: 2003040141
Program Studi	: Teknik Informatika
Fakultas	: Teknik dan Sains
Perguruan Tinggi	: Universitas Muhammadiyah Purwokerto
Judul	: Analisis Sentimen Terhadap Rencana Kenaikan Pajak Pertambahan Nilai Menjadi Duabelas Persen Pada Media Sosial X Dengan Algoritma *Support Vector Machine*

telah disetujui untuk diajukan dalam seminar proposal

Purwokerto, …….

Pembimbing

Sigit Sugiyanto, S.T., M.Eng.

NIK. 2160766

BAB I
PENDAHULUAN

A. Latar Belakang Masalah

Pajak merupakan elemen krusial dalam mendukung pembangunan nasional dan menjadi sumber keuangan strategis utama bagi negara maka pajak merupakan sumber pendapatan negara untuk membiayai semua pengeluaran pemerintahan seperti pembelanjaan pegawai, pembelian barang, pembangunan, pemeliharaan insfrastruktur, dan lain sebagainya. Pajak berfungsi sebagai penghimpunan dana dari masyarakat ke dalam kas negara yang diperuntukan bagi pembiyaan pengeluaran pemerintah. Undang-undang Republik Indonesia nomer 28 tahun 1983 tentang perubahan ketiga atas undang-undang nomer 6 tahun 1983 tentang ketentuan umum dan tata cara perpajakan pasal 1 ayat 1 menjelaskan bahwa pajak merupakan konstribusi wajib kepada negara yang terutang oleh orang pribadi atau badan bersifat memaksa berdasarkan undang-undang dengan tidak mendapatkan imbalan secara langsung digunakan untuk keperluan negara bagi sebesar-besarnya kemakmuran rakyat.

Pajak pertambahan nilai (PPN) merupakan pajak atas pajak atas komsumsi barang dan jasa di dalam daerah peabean yang dikenakan secara bertingkat dalam setiap jalur produksi dan distribusi. PPN merupakan pajak tidak langsung karena pembayaran atau pemungutan pajak disetorkan oleh pihak lain yang bukan

penanggung pajak. Negara Indonesia mulai menerapkan pajak pertambahan nilai (PPN) pada tahun 1983 setelah sebelumnya menggunakan sistem pajak penjualan (PPN) pada tahun 1951. Peralihan sistem ini didasari oleh penerapan pungutan sebelumnya yang tidak dapat memnuhi aktivitas Masyarakat yang terus berkembang. Pajak pertambahan nilai (PPN) dikenakan atas setiap penyerahan barang kena pajak (BPK) atau jasa kena pajak (JKP) menurut ketentuan perundang-undangan. Namun, untuk tujuan tertentu, pemerintah memberi fasilitas untuk tidak mengenakan PPN atas jenis barang, jasa dan sektor usaha tertentu. Fasilitas PPN merupakan bentuk-bentuk perlakuan khusus terkait pungutan pajak pertambahan nilai (PPN) atas barang atau kegiatan tertentu. Dalam memaksimalkan kewajiban atas pajak pertambahan nilai suatu entitas diperlukan pengetahuan terkait PPN dan fasilitasnya.

Pada tahun 2022, pemerintahan menaikan tarif PPN dari 10% menjadi 11% dengan tujuan untuk mencapai target penerimaan pajak dan mengatasi dampak ekonomi dan meningkatkan pendapatan negara, disisi lain para pengusaha dan pedagang khawatir bahwa daya beli masyarakat akan menurun jika tarif PPN terus meningkatkan hingga 12% paling lambat pada tahunn 2025. Rencana kenaikan tarif PPN menjadi 12% berpotensi menimbulkan permasalahan di Tengah Masyarakat, karena dampaknya tidak hanya dirasakan oleh kelompok tertentu, tetapi berpengaruh pada seluru lapisan Masyarakat, baik sebagai pengusaha maupun sebagai konsumen akhir. Pengusaha menghadapi beban yang cukup berat

akibat selisih antara tarif PPN yang berlaku saat ini dan tarif sebelumnya, tertentu dalam transaksi dengan jumlah besar. Selain itu, pengusaha harus menyesuaikan harga jual barang atau jasa, yang pada akhirnya berdampak pada konsumen. Konsumen mempertimbangkan harga produk dan jasa berbagai faktor utama dalam keputusan pembelian mereka.

Masyarakat Indonesia menyampaikan berbagai tanggapan terhadap rencana pemerintah untuk menaikan tarif pajak pertambahan nilai (PPN) menjadi 12% melalui media sosial X. Pemerintah menetapkan kebijakan ini dalam undang-undang harmonisasi peraturan perpajakan (UU HPP) sebagai Upaya untuk meningkatkan penerimaan negara dan menyeimbangkan anggaran pendapatan dan belanja negara. Kenaikan tarif ini menimbulkan beragam respon dari Masyarakat, baik yang mendukung maupun yang menolak. Masyarakat menganggap kebijakan ini sebagai Langkah strategis pemerintah dalam memperkuat ekonomi nasional. Sebagaian lainya menilai bahwa kenaikan tarif PPN akan memberatkan Masyarakat karena dapat menurunkan daya beli dan menaikan harga barang serta jasa. Pengguna media sosial X mengunggah opini mereka dalam bentuk komentar, cuitan, maupun diskusi publik yang mewakili presepsi Masyarakat terhadap kebijakan fisikal tersebut. Media sosial X berfungsi sebagai sasaran partisipasi publik yang memungkinkan Masyarakat untuk menyuarakan pendapat dan mengkritik kebijakan pemerintah secara terbuka. Keberagaman tanggapan masyarakat menunjukan perlunya pemerintah untuk memahami opini publik

secara mendalam. Analisis sentiment diperlukan untuk mengidentifikasi presepsi masyarakat secara sistematis, sehingga pemerintah dapat merumuskan kebijakan yang lebih responsive dan sesuai dengan kondisi sosial ekonomi masyarakat.

Support Vector Machine (SVM) merupakan salah satu metode pembelajaran yang digunakan untuk menyelesaikan permasalahan klasifikasi dan regresi, termasuk dalam bidang analisis sentimen. Dengan manfaatkan algoritma SVM untuk mengklasifikasi opini masyarakat berdasarkan pola-pola yang terdapat dalam teks. Pengguna menerapkan SVM menggunakan bahasa pemograman Python pada platfrom Google Colab. Platfrom tersebut memberikan dukungan dalam proses data secara efesien. Kajian ini menggunakan penerapan SVM melalui dua tahapan utama, yaitu tahap pelatihan dan tahap pengujian. Pada tahap pelatihan, menggunakan model berdasarkan data historis, selanjutnya, pada tahap pemgujian, peneliti menggunakan data berlabel untuk menyelesaikan tugas klasifikasi secara sepesifik.

B. Peruusan Masalah

Berlandaskan terhadap pembahasan latar belakang permasalahan, berikut adalah permasalaham yang telah dirumuskan:

Rencana kenaikan tarif Pajak Pertambahan Nilai (PPN) menjadi 12% yang telah diatur dalam Undang-Undang Harmonisasi peraturan Perpajakan (UU HPP)

menimbulkan berbagai respons di kalangan masyarakat, khususnya di media sosial X. Kebijakan ini berpotensi memengaruhi daya beli masyarakat, sektor usaha, serta stabilitas ekonomi secara keseluruhan. Oleh karena itu, analisis sentimen diperlukan untuk memhami opini public terhadap kebijakan tersebut. Dalam penelitian ini, algoritma Support Vector Machine (SVM) digunakan untuk mengklasifikasikan sentimen masyarakat berdasarkan data dari media sosial X. Diharapkan hasil analisis ini dapat memberikan wawasan yang berguna bagi pemerintah dalam merumuskan kebijakan yang lebih responsive serta meningkatkan pemahaman masyarakat terhadap kebijakan perpajakan.

C. Batasan Masalah

Ruang lingkup Batasan masalah antara lain:

1. Kajian ini menggunakan dua label sentimen, yaitu label positif dan negatif. Data yang digunakan dalam penelitian ini berasal dari platform X dengan rentang waktu pengambilan data dari tahun 2024. Pengumpulan data dilakukan dengan menggunakan kata kunci seperti *"kenaikan ppn 12%"* dan *"tolak kenaikan PPN 12%"* untuk mendapatkan opini publik yang relevan. Data yang diambil difokuskan pada teks berbahasa Indonesia untuk memastikan analisis lebih sesuai dengan konteks sosial di Indonesia.
2. Fokus analisis akan terbatas pada teks-teks tertulis yang terdapat dalam posting atau komentar pengguna media sosial terkait topik kenaikan PPN

menjadi 12%. Aspek-aspek lain dari media sosial seperti gambar, *video*, atau *audio* tidak akan menjadi bagian dari analisis dalam penelitian ini.

D. Tujuan Penelitian

Tujuan dari penelitian ini antara lain:

1. Mengidentifikasi dan menganalisis sentiment positif dan negatif pada media sosial X terkait kenaiakan PPN menjadi 12%.
2. Menggunakan analisis sentimen untuk mengevaluasi bagaimana masyarakat merespons kebijakan kenaikan PPN menjadi 12% yang diterapkan, serta untuk memahami dukungan atau ketidaksetujuan terhadap kebijakan tersebut.
3. Menguji keefektifan metode Support Vector Machine (SVM) dalam mengklasifikasikan sentimen dari postingan terkait kenaikan PPN menjadi 12%.

E. Manfaat Penelitian

Manfaat yang diperoleh dalam penelitian ini adalah:

1. Kajian ini akan memberikan imemberikan gambaran mengenai respons masyarakat terhadap kebijakan kenaikan PPN, yang dapat menjadi bahan evaluasi bagi pemerintah untuk memastikan bahwa kebijakan ini tidak hanya fokus pada optimalisasi pendapatan negara, tetapi juga mempertimbangkan dampaknya terhadap daya beli masyarakat.
2. Mengevaluasi lebih akurat terhadap efektivitas kebijakan kenaiakan PPN menjadi 12% seperti penerimaan masyarakat terkait kebijakan tersebut serta kontroversi yang muncul.

BAB II
TINJAUAN PUSTAKA

A. Penelitian Terdahulu

Penelitian-penelitian terdahulu yang telah dilakukan oleh beberapa peneliti antara lain:

1. Penelitian yang dilakukan oleh (Jesica Krisrovina Siagian & Painem, 2024) tentang analisis sentiment Masyarakat Indonesia terhadap rencana kenaikan PPN menjadi 12% di media sosial X dengan menggunakan metode Naïve bayes. Studi ini menggunakan 468 dataset yang diperoleh melalui proses crawling data di Twitter untuk menganalisis sentiment Masyarakat Indonesia terkait kenaikan PPN menjadi 12%. 326 data (77,3%) bersentimen negatif dan 106 data (22,7%) bersentimen positif sejak 1 maret 2024- 15 mie 2024. Penenlitian ini melibatkan enam tahap uatama, yaitu proses crawling data, pemberian label (labeling), preprocessing, pembagian data, ektraksi fitur menggunakan bag clasiffer, serta pengujian menggunakan confusion matrix. Dari hasil pengujian dan evaluasi, diperoleh akurasi sebesar 83%, recall sebesar 78,6%, dan presisi sebesar 68,8%.
2. Penelitian yang dilakukan oleh (Novi Fauziah & rekan-rekanya, 2024) yang berjudul pelabelan vader dalam menganalisis presepsi masyarakat terhadap kenaikan tarif PPN di indonesia. Sentiment yang dianalisis dari

media sosial Tritter menunjukan bahwa sentiment Masyarakat dominan negatif karena banyaknya kekhawatiran serta kritikan masyarakat terkait kenaikan tarif PPN yang dirasa hanya membebani masyarakat kecil-menengah. Terhadap pula sentimen positif yang berisi dukungan serta optimisme terhadap kebijakan kenaikan tarif PPN untuk percepatan pembangunan ekonomi nasional. Adanya dominasi sentimen negatif terhadap kenaikan tarif PPN menunjukkan bahwa pentingnya bagi pemerintah untuk melihat dampak ke depan dari kenaikan tarif PPN serta survei mendalam dengan mempertimbangkan kritikan dan saran masyarakat dalam penentuan kebijakan selanjutnya. Selain itu dialog dan transparansi dari pemerintah sangat penting untuk membangun kepercayaan publik dan memastikan bahwa kenaikan PPN digunakan untuk kepentingan rakyat. Berdasarkan dari data dari 2.071 data tweets yang telah diolah, sentimen negatif sebesar 78%, sentimen positif memiliki persentase 6%, dan sentimen netral sebesar 16%. Sentimen negatif memiliki persentase yang lebih besar karena banyaknya kritikkan masyarakat yang tidak setuju.

3. Penelitian yang dilakukan oleh (Rahadi Rahma & teman-temannya, 2023) yang berjudul analisis sentiment pengguna Twitter menggunakan support vector machine pada kasus kenaikan BBM. Metode yang digunakan dalam analisis sentimen adalah Support Vector Machine (SVM), yang menganalisis komentar masyarakat di Twitter terkait kenaikan harga BBM. Penelitian ini memanfaatkan 258 data komentar yang diambil pada

4 September 2022, tepat sehari setelah kenaikan harga BBM. Tahap awal penelitian mencakup preprocessing untuk menghapus kata-kata atau informasi yang tidak relevan. Selanjutnya, data dibagi menjadi 80% untuk pelatihan (training) dan 20% untuk pengujian (testing). Hasil pengujian menunjukkan tingkat akurasi sebesar 82,69%, spesifisitas 79,07%, sensitivitas 100%. Dari 52 data yang diuji, terdapat 9 komentar positif dan 43 komentar negatif, sehingga disimpulkan bahwa mayoritas masyarakat tidak setuju dengan kenaikan harga BBM.

4. Studi yang dilakukan oleh (Zidan Alhaq & koleganya) Penelitian ini membahas penerapan metode *Support Vector Machine* (SVM) untuk analisis sentimen pengguna Twitter, dengan fokus pada topik yang sering diperbincangkan, yaitu marketplace. Bukalapak, sebagai salah satu marketplace terpopuler di Indonesia, menyediakan layanan transaksi yang cepat dan aman bagi penggunanya. Ulasan dari pengguna dapat berupa sentimen positif, negatif, atau netral. Oleh karena itu, diperlukan metode yang mampu mengidentifikasi opini pengguna Bukalapak di media sosial Twitter. Untuk menyelesaikan permasalahan ini, data yang diperoleh dari Twitter dilabeli dan dianalisis menggunakan metode SVM untuk mengelompokkan opini-opini tersebut. Hasil klasifikasi dengan SVM menunjukkan tingkat akurasi sebesar 93%.
5. Penelitian yang dilakukan oleh (Yuris Alkhalidi, Windu Gata, Arfhan Prastyeo, dan Imam Budiawan, 2020) mengenai analisis sentimen penghapusan ujian nasional pada twitter menggunakan *support vector*

machine dan *naïve bayes* berbasis particle swarm optimization. Dalam penggunanya twitter digunakan sebagai platfromyang membahas tentang opini public, hiburan dan tranding topik didunia salah satu perbincangan pada awal tahun 2020 yakni dihapusnya ujian nasional (UN) oleh kementrian Pendidikan dan kebudayaan republic Indonesia (mendikbud RI). Opini dan sentiment pengguna di twitter pun sangat beragam, ada yang termasuk kedalam sentiment positif dan ada juga sentiment negatif. Untuk memilah mana yang termasuk ke dalam sentiment positif dan negatif diperlukan sebauh rangkaian proses, salah satu proses yang dapat digunakan yakni data mining. Pengujian dilakukan menggunakan *k-fold cros validation* untuk diperoleh nilai akurasi (*accuracy)*, tabel *confusion matrix* dan *area under curve.* Hasil pengujian diperoleh nilai akurasi 92,92% dan ACU sebesar 0,977 untuk SVM tanpa PSO. Lalu nilai akurasi 94,81% dan ACU sebesar 0,974 untuk SVM dengan PSO. Nilai akurasi 85,93% dan ACU sebesar 0,645 untuk NB tanpa PSO. Serta nilai akurasi 86,92% dan ACU sebesar 0,715.

6. Penelitian yang dilakukan oleh (Hendry Cipta Husada dan Adi Suryaputra Paramita) dengan judul analisis sentiment pada maskapai penerbangan di platfrom twitter menggunakan algoritma *support vector machine*. Perkembangan teknologi saat ini telah memberikan kemudahan bagi banyak orang dalam mendapatkan dan menyebarkan informasi di berbagai *social media platfrom*. Twitter merupakan salah satu media yang kerap digunakan untuk menyampaikan opini sebagai bentuk reaksi seseorang

atas satu hal. Proses analisis sentiment dilakukan dengan proses data *preprocessing*, pembobotan kata menggunakan metode TF-IDF, penerapan algortima, dan pembahasan atas klasifikasi. Klasifikasi opini dilakukan dengan *machine learning approach* memanfatkan algortima*multi-class support vector machine* (SVM). Data yang digunakan dalam penelitian ini adalah opini dalam bahasa inggris dari pengguna twitter terhadap maskapai penerbangan. Berdasarkan pengujian yang telah dilakukan, hasil klasifikasi terbaik diperoleh menggunakan SVM karne RBF pada nilai parameter C(*complexity)* = 10 dan y(*gamma)* = 1, dengan nilai *accuracy* sebesar 84,37% dan 80,41% Ketika menggunakan 10-*fold cros validation.*

7. Penelitian yang dilakuakan oleh (Huang, 2023) dengan judul *Sentiment analysis for social media using SVM classifier of machine learning.* Penelitian ini bertujuan untuk mengetahui seberapa baik kinerja *Support Vector Machine* (SVM) ketika diberi tugas menganalisis perasaan orang, Untuk mengevaluasi seberapa baik SVM bekerja,kami menggunakan satu kumpulan data pra-klasifikasi yang berasal dari tweet. Hasil dari penelitian ini adalah Metrik presisi, recall, dan f-measure digunakan untuk melakukananalisis akurasi pada hasil. Menurut penyelidikan, kumpulan data tersebut memiliki akurasi 91,8persen, presisi 91,3 persen, dan recall 82,8 persen. Selain itu, nilai f1 bisadinyatakan sebesar 86,9. Keterbatasan penelitian ini yaitu pada penelitian ini tidak ada proses normalisasi kata, sehingga masih ada slang word pada hasil preprocessing.

8. Penelitian yang dilakukan oleh (Aditiya Hermawan, Indrico Jowense, Junaedi, dan Edy , 2023) yang berjudul implementasi tex-mining untuk analisis sentiment pada twitter dengan algoritma *support vector machine.* Setiap tahun, jumlah orang yang menggunakan media sosial bertambah seiring dengan jumlah orang yang menggunakan internet. Peningkatan tersebut diiringin dengan meningkatnya informasi pada internet yang tentunya informasi tersebut mempunyai nilai jika dilakukan analisis. Untuk menganalisis data dalam jumlah besar dapat menggunakan Teknik text mining. Text mining mampu memproses untuk memperoleh informasi berkualitas tinggi dari teks. Penggunaan text mining menggunakan SVM dalam melakukan klasifikasi pada tweet berbahasa Indonesia mempunyai akurasi 73% berdasarkan pada 10 kali percobaan yang dilakukan dengan *keyword* dan waktu yang berbeda-beda. Kemudian nilai presisi yang didapatkan adalah 67% dan nilai *recall* yang didapatkan 54%.

Tabel 2.1 Penelitian Terdahulu

No	Peneliti	Judul	Metode	Hasil	Keterbatasan
1.	(Jesica Kristoviani Siagian dan Painem, 2024)	*Analisis sentiment Masyarakat Indonesia terhadap rencana kenaikan PPN menjadi 12% di media sosial X dengan menggunakan metode naïve bayes.*	*Naïve bayes*	Hasil yang diperoleh 468 dataset yang diperoleh melalui proses crawling data di twitter untuk menganalisis sentiment Masyarakat di Indonesia terkait rencana kenaikan PPN 12%. 326 data (77,3%) bersentimen negatif 106 data (22,7%) bersentimen positif sejak 1 maret 2024-15 mie 2024. Dari hasil pengujian diperoleh akurasi sebesar 83%, recall sebesar 78,6% dan presisi sebesar 68,8%.	Keterbatasan pada penelitian ini yaitu tidak terdapat proses normalisasi kata, yang dimana data yang berisikan kalimat tidak baku tidak diubah.
2.	(Novi Fauziah dan rekan-rekanya 2024)	*pelabelan vader dalam menganalisis presepsi Masyarakat terhadap kenaikan tarif PPN di indonesia*	*Support Vector Machine*	Hasil penelitan dari media sosial twitter menunjukkan sentimen positif yang berisi dukungan serta optimisme terhadap kebijakan kenaikan tarif PPN untuk percepatan pembangunan ekonomi nasional. Berdasarkan dari data dari 2.071 data tweets yang telah diolah, sentimen negatif sebesar 78%, sentimen positif memiliki persentase 6%, dan sentimen netral sebesar 16%. Sentimen negatif memiliki persentase yang lebih besar karena banyaknya kritikkan masyarakat yang tidak setuju	Keterbatasan pada penelitian ini yaitu penggunaan metode perluasan akronim, bahasa gaul penerjemahan kata, dan penerjemahan emoji pada tahap preprocessing.
3.	(Rahadi Rahma dan teman - temanya, 2023)	*Analisis sentiment pengguna twitter menggunakan support vector machine pada kasuk kenaikan BBM*	*Support Vector Machine*	Penelitian ini memanfaatkan 258 data komentar yang diambil pada 4 September 2022, tepat sehari setelah kenaikan harga BBM. Tahap awal penelitian mencakup *preprocessing* untuk menghapus kata-kata atau informasi yang tidak relevan. Selanjutnya, data dibagi menjadi 80% untuk pelatihan (*training*) dan 20% untuk pengujian (testing). Hasil pengujian menunjukkan tingkat akurasi sebesar 82,69%, spesifisitas 79,07%, sensitivitas 100%. Dari 52 data yang diuji, terdapat 9 komentar positif dan 43 komentar negatif.	Keterbatasan pada penelitian ini yaitu teknik TF-IDF digunakan untuk mengubah kalimat-kalimat abstrak menjadi vektor sehingga menjadi dimodelkan dengan SVM.

4.	(Zidan Alhaq dan koleganya)	Penerapan metode *support vector machine* (SVM) untuk analisis sentiment pengguna	SMOTE dan SVM	Hasil dari penelitian ini adalah klasifikasi sentimen masyarakat terhadap pemblokiran situs judi online dapat dilakukan dengan mengunakan metode SVM dengan pembobotan TFIDF dan penyetaraan data SMOTE. Klasifikasi sentimen dari teks komentar YouTube mencapai nilai akurasi sebesar 61.84% dan mencapai nilai F1-score 0.7590	Keterbatasan pada penelitian ini yaitu pada penelitian ini tidak dijelaskan mengenai sistem informasi, namun penelitian ini lebih menjelaskan tahap preprosesing
5.	(Yusri Alkhalidi, Windu Gata, Arfhan Prastyo, dan Imam Budiawan, 2020)	analisis sentimen penghapusan ujian nasional pada twitter menggunakan support vector machine dan naïve bayes berbasis particle swarm optimization	*Support Vector Machine*	Hasil Untuk memilah mana yang termasuk ke dalam sentiment positif dan negatif diperlukan sebauh rangkaian proses, salah satu proses yang dapat digunakan yakni data mining. Pengujian dilakukan menggunakan k-fold cros validation untuk diperoleh nilai akurasi (accuracy), tabel confusion matrix dan area under curve. Hasil pengujian diperoleh nilai akurasi 92,92% dan ACU sebesar 0,977 untuk SVM tanpa PSO. Lalu nilai akurasi 94,81% dan ACU sebesar 0,974 untuk SVM dengan PSO. Nilai akurasi 85,93% dan ACU sebesar 0,645 untuk NB tanpa PSO. Serta nilai akurasi 86,92% dan ACU sebesar 0,715.	Keterbatasan penelitian ini yaitu pada tahap pre-processing tidak terdapat proses cleaning dan normalisasi.
6.	(Hendry Cipta Husada dan Adi Syahputra Pramita)	analisis sentiment pada maskapai penerbangan di platfrom twitter menggunakan algoritma support vector machine	*Support vector machine*	Hasil klasifikasi opini dilakukan dengan machine learning approach memanfatkan algortimamulti-class support vector machine (SVM). Data yang digunakan dalam penelitian ini adalah opini dalam bahasa inggris dari pengguna twitter terhadap maskapai penerbangan. Berdasarkan pengujian yang telah dilakukan, hasil klasifikasi terbaik diperoleh menggunakan SVM karne RBF pada nilai parameter C(complexity) = 10 dan y(gamma) = 1, dengan nilai accuracy sebesar 84,37% dan	Keterbatasan pada penelitian ini yaitu ketidakcukupan data, keakuratan prediksi pengklasifikasi dengan metode penyematan Word2Vec rendah sehingga makalah ini hanya mengumpulkan opini masyarakat Inggris di Twitter tentang maskapai penerbangan.

					80,41% Ketika menggunakan 10-fold cros validation.	
7.	(Huang, 2023)	*Sentiment analysis for social media using SVM classifier of machine learning*	*Support*	*Vector Machine*	Hasil dari penelitian ini adalah Metrik presisi, recall, dan f-measure digunakan untuk melakukananalisis akurasi pada hasil. Menurut penyelidikan, kumpulan data tersebut memiliki akurasi 91,8persen, presisi 91,3 persen, dan recall 82,8 persen. Selain itu, nilai f1 bisadinyatakan sebesar 86,9	Keterbatasan penelitian ini yaitu pada penelitian ini tidak ada proses normalisasi kata, sehingga masih ada slang word pada hasil preprocessing.
8.	(Aditiya Hermawan, Indrico Jowenes, Junaedi, dan Edy, 2023)	implementasi tex-mining untuk analisis sentiment pada twitter dengan algoritma support vector machine.	*Support*	*Vector Machine*	ext mining mampu memproses untuk memperoleh informasi berkualitas tinggi dari teks. Penggunaan text mining menggunakan SVM dalam melakukan klasifikasi pada tweet berbahasa Indonesia mempunyai akurasi 73% berdasarkan pada 10 kali percobaan yang dilakukan dengan keyword dan waktu yang berbeda-beda. Kemudian nilai presisi yang didapatkan adalah 67% dan nilai recall yang didapatkan 54%.	Keterbatasan penelitian ini yaitu pada penelitian ini tidak ada proses normalisasi kata, sehingga masih ada slang word pada hasil preprocessing.

B. Landasan Teori

1. *Crawling*

Crawling adalah prosesotomatis untuk mengumpulkan dan mengideks data dari berbagai sumber seperti situs web, database, atau dokumen. Proses ini menggunakan perangkat lunak khusus yang disebut “crawler” atau “bot” untuk mengakses sumber data dan mengambil informasi yang dibutuhkan. Data yang dikumpulkan melalui crawling kemudian dapat diperoses dan digunakan untuk berbagai tujuan, seperti analisis data, penelitian, atau pengembangan sistem informasi. Proses crawling data dimulai dengan crawler yang menjelajahi internet dan mengideks serta mengumpulkan data dari berbagai sumber(Alhaq et al., n.d.-a). Data yang dikumpulkan dapat digunakan sebagai alat untuk pengembangan sistem atau sebagai data yang biasanya digunakan oleh mesin pencarian untuk menampilakan hasil pencarian yang lebih relevan. Tujuan dari crawling data adalah

a. Mengumpulkan data besar dari berbagai sumber seperti situs web, database, atau dokumen dalam waktu singkat dan efesien.
b. Menggunakan data yang dikumpulkan untuk melakukan analisis data seperti analisis pasar, analisis prilaku pelanggan, dan lain-lain.
c. Menggunakan data yang dikumpukan untuk melakukan penelitian seperti penelitian pasar, penenlitian sosial dan lain-lain.

d. Membuat database yang mengandung informasi dari berbagai sumber seperti situs web, database, atau dokumen.
e. Memantau informasi dari berbagai sumber seperti media sosial, situs web, dan lain-lain untuk memastikan infromasi yang diterima akurat dan terkini.
f. Menggunkan daya dikumpilkan untuk membangun aplikasi seperti aplikasi pencarian, aplikasi e-commerce, dan lan-lain.

2. Analisis Sentimen

Analisis sentimen adalah sutdi komputasi yang bertujuan untuk memahamiopini, sikap, dan emosi sesorang terhadap suatu topik tertentu. Hasil analisis ini biasanya diklasifikasikan sebagai sentiment positif atau negatif. Analisis sentiment melibatkan proses pendekteksian polaritas teks untuk menentukan apakah teks tersebut naegatif, positif, atau netral.

3. Media Sosial X

Sosial media adalah tempat yang digunakan orang-orang untuk mengeluarkan pendapat mereka tentang berbagai topik. Pemakai sosial media di Indonesia sangat besar, hal ini mendorong munculnya data teksual yang tidak terbatas. Salah satunya pemanfaatan data ini adalah mengetahui sntimen public tentang kenaikan PPN 12%(Mega Putri, 2024).

X adalah salah satu platform media sosial yang terkenal di kalangan masyarakat umum, termasuk di Indonesia dan di seluruh dunia, serta populer di kalangan pelajar. Platform ini menghubungkan pengguna dengan informasi mengenai topik-topik yang relevan. *Twitter* muncul setelah kepopuleran *Facebook* sebagai platform media sosial yang mengusung konsep *microblogging*, di mana setiap *tweet* atau cuit memiliki batasan 280 karakter. Awalnya, batasan karakter untuk setiap *tweet* adalah 140 karakter, namun jumlah ini ditingkatkan seiring berjalannya waktu. Perubahan ini mempermudah pengguna untuk mengumpulkan informasi dengan lebih efisien (Pamungkas et al., n.d.)

Beberapa istilah yang umum digunakan di platform Twitter "X" antara lain sebagai berikut:

a. *Tweet*: Pesan atau status yang ditulis dalam kotak shout yang dapat berisi informasi, gambar, opini, dan rangkaian pesan lainnya. Tweet memiliki batasan jumlah karakter, yaitu 280 huruf.

b. *Mention*: Digunakan untuk menandai atau memanggil pengguna Twitter lain dalam sebuah tweet dengan menambahkan "@" diikuti dengan nama pengguna yang dimaksud.

c. *Reply*: Tanggapan atau balasan terhadap tweet dari pengguna lain.

d. *Retweet*: Biasa disingkat sebagai RT, digunakan untuk menunjukkan setuju dengan isi dari tweet pengguna lain dan membagikannya ke pengikut kita.

e. *Like*: Digunakan untuk menunjukkan bahwa pengguna menyukai tweet yang diposting oleh pengguna lain.

f. *Direct Message*: Biasa disingkat sebagai DM, merupakan fitur untuk mengirim pesan secara pribadi kepada pengguna lain tanpa diketahui oleh pengikut kita.

g. *Hashtag*: Digunakan untuk menandai sebuah topik atau tema dalam sebuah tweet dengan menggunakan tanda "#" diikuti dengan kata kunci yang relevan. Hashtag membantu meningkatkan visibilitas tweet.

h. *Trending Topic*: Topik atau tema yang sedang populer atau banyak dibicarakan oleh pengguna Twitter karena mendapatkan perhatian yang signifikan.

4. *Google Colaboratory*

Google Colaboratory, yang dikenal sebagai *Google Colab*, adalah sebuah tool yang digunakan untuk tujuan penelitian secara gratis dan menggnukana cloud atau sistem penyimpanan awan. *Google Collaboratory* dibangun dengan menggunakan elemen dari jupyter serta penelitian salah satunya adalah datamining(Hakim, 2021). *Google collaboratory* pada

dasaarnya mempunyai kesamaan fungsi dengan jupyter Notebook, letak perbedaanya adalah *Google Collaboratory* dapat di akses secara online serta gratis. Beberapa fitur utama dari google colab adalah:

a. *Python* di *cloud*: Anda dapat menulis dan mengeksekusi kode pyhton langsung di browser web anda tanpa perlu menginstal pyhton atau Pustaka di computer local.
b. Gratis: Google colab adalah layanan gratis yang disediakan oleh Google. Anda dapat menggunakannya tanpa biaya.
c. GPU gratis: Google colab menyediakan akses ke GPU (*Graphics Processing Unit)* secara gratis. Ini sangat berguna untuk pelatihan model mesin yang memerlukan daya komputasi yang sangat tinggi.
d. Akses ke penyimpanan Google Drive: anda dapat mengakses dan menyimpan file langsung di Google Drive anda, yang memudahkan berbagai dan menyimpan pekerjaan anda.
e. Notebook Interaktif: Google Colab menggunakan format “notebook” yang memungkinkan anda untuk mengabungkan kode, teks, gambar, dan hasil dalam satu dokumen interaktif. Ini sangat berguna untuk dokumentasi dan berbagai hasil analisis.
f. Pustaka Tersedia: Google Colab menyediakan banyak Pustaka umum seperti NumPy, pandas, TensorFlow, PyTorch, dan sebagainya. Anda dapat mengimpor Pustaka-pustaka ini kedalam lingkungan colab dengan mudah.

g. Kerja sama tim: anda dapat berbagai notebook colab dengan anggota tim anda dan berkolaborasi dalam waktu nyata.
h. Fleksibel dan mudah digunakan: meskipun kuat, Google Colab adalah alat yang ramah untuk pemula. Ini adalah cara yang baik untuk memulai pemograman Python dan eksplorasi data tanpa kerumitan konfigurasi lokal.

5. *Python*

Python adalah bahasa pemrograman tingkat tinggi yang diciptakan oleh Guido van Rossum dan pertama kali dirilis pada tahun 1991. *Python* dikenal karena sintaksnya yang mudah dipahami dan bersih, menjadikannya pilihan populer bagi pemula dan pengembang berpengalaman(Kristovani Siagian, 2024b). *Python* dilengkapi dengan pustaka standar yang sangat luas serta ekosistem pustaka pihak ketiga yang kaya, seperti *NumPy*, *Pandas*, *Matplotlib*, *TensorFlow*, scipy, scikit learn, Theano, keras, pytroch, yang memudahkan pengembangan aplikasi di berbagai bidang seperti data *science*, *machine learning*, dan web *development*.

Menurut (*Buku Python*, n.d.) adalah bahasa pemograman tingkat tinggi yang bersifat interpreter, interaktif, objek oriented dan dapat berjalan di hamper semua platfrom. Python sebagai tingkat tinggi yang mudah untuk dipelajari karena sintaknya yang jelas dan juga elegan, karena sintaknya lebih menggunakan bahasa manusia daipaada bahasa computer, dan memiliki modul-modul yang edesien dan siap langsung digunakan. Scoure code bahasa

python akan dikomplikasi menjadi format bytecode yang dieksekusi. Kode python lebih kambat dieksekusi dibandingkan dengan bahasa program lain yang bersifat low-level (sivam,2018). Keunggulan bahasa program python menurut (kadarina and ibnu fajar, 2019):

a. Merupakan bahasa program tingkat tinggi yang mudah dipelajari karena sintaknya yang jelas dan mudah dibaca karena lebih mendekati bahasa manusia.
b. Tersedia banyak library yang dapat digunakan, yang kebanyak ditulis oleh bahasa C.
c. Bahasa pyhton dapat berjalan diberbagai platfrom tanpa harus menulis kode untuk platfrom tertentu.
d. Dapat digunakan untuk mengembangkan berbagai hal seperti software, hardware, internet of things, web development, video game, dan mobile apps.

6. Metode *Support Vector Machine* (SVM)

Support Vector Machine (SVM) adalah pemodelan data empiris dapat menimbulkan beberapa permasalahan Ketika data yang dipel=roleh berdemensi tinggi (ruang fitur) dan tidak seragam yang bisa melibatkan analisis dengan pendekatan *Nerual Network* (NN) tradisional mengalammi kesulitan dalam generalisasi dan menghasilkan model yang bisa *overfit* data. SVM dikembangkan untuk memecahkan masalah klasifikasi kerena SVM

memiliki kemampuan yang lebih baik dalam mengeneralisasi data bila dibandingkan dengan Teknik yang sudah ada sebelumnya.

SVM merupakan sistem pembelajaran menggunakan ruang berupa fungsi-fungsi linear dalam sebuah ruangan berdimensi tinggi yang dilatih menggunakan algortima pembelajaran berdasarkan pada teori optimasi dengan mengimplementasikan *learning bias*(Husada & Paramita, 2021). Pendekatan dengen menggunakan SVM ini memiliki banyak manfaat lain seperti misalnya model yang dibangun memiliki ketergantungan ekplisit pada subset dari datapoints, serta *support vector* yang membantu dalam interprestasi.

Kelebihan SVM diantaranya efektif dalam menangani data dengan dimensi tinggi, seperti teks atau gambar, SVM hanya bergantung pada *support vectors*, sehingga membutuhkan ruang memori yang relatif kecil, melalui penggunaan kernel, SVM dapat memisahkan data yang tidak linier secara efektif(Samsudiney, 2019).

Persamaan *Support Vector Machine* (SVM) dapat dilihat pada persamaan 2.1.

$$f(x) = w.x + b \qquad (2.1)$$

Keterangan :

w : Parameter yang dicari (garis yang tegak lurus antar garis dan titik *support vector*)

x : Titik data masukan *Support Vector Machine*

b : Parameter yang dicari (nilai bias)

Atau

$$f(x) = \sum_{i=1}^{M}(a_i\, y_i\, K\,(X_I, X_I) + b) \qquad (2.2)$$

Keterangan:

$aiyi$: Nilai bobot setiap titik data

$K\,(x, xi)$: Fungsi kernel

b : Parameter hyperlane yang dicari (nilai bias)

Penelitian ini menggunakan karnel linear. Persamaan yaitu:

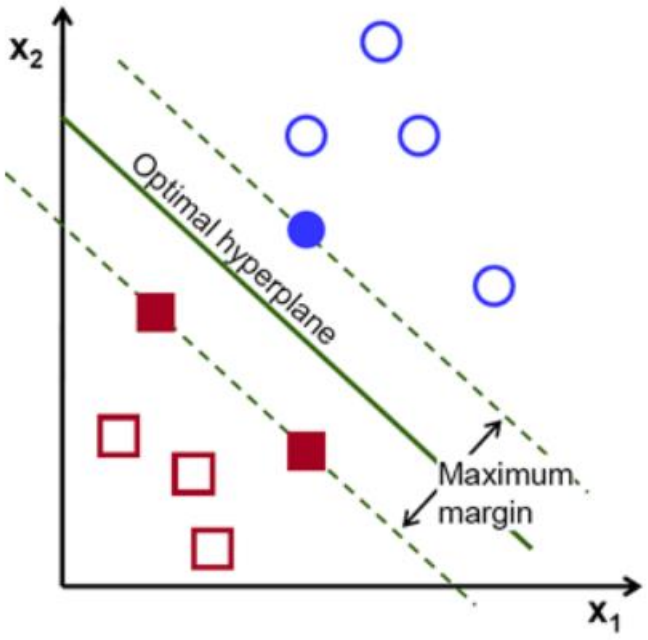
$$f(x) = w.x + b$$

$K\,(x, y)$: Nilai kernel dari data x dan y

x : Fitur data 1

y : Fitur data 2

Ilustrasi gambar metode *support vector machine* dijelaskan pada Gambar 2.1



Gambar 2. 1 Ilustrasi metode *support vector machine*

7. *Term Frequency-Inverse Document Frequency* (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) adalah metode yang digunakan untuk menentukan nilai frekuensi sebuah kata di dalam sebuah dokumen atau artikel dan juga frekuensi di dalam banyak dokumen. Perhitungan ini menentukan seberapa relevan sebuah kata di dalam sebuah dokumen (evan, 2014). TFIDF adalah sebuah algortima yang umumnya digunakan untuk mengolah data besar(Kamath, 2014). Algoritma TF-IDF melakukan pemberian bobot pada setiap kata kunci disetiap kategori untuk mencari kemiripan kata kunci dengan kategori yang tersedia.

Sebelum melakukan pembobotan maka akan dilakukan lima tahap pencarian text preprocessing yaitu pemecah kalimat, case folding, tokenizing, filtering, dan stemming, lalu selanjutnya dilakukan proses menghitung bobot TF-IDF, bobot quert relevance dan bobot similarity (Alhaq et al., n.d.-b). TF-IDF pada dasarnya merupakan hasil dari perhitungan antara TF (Term Frequency) dan IDF (Inverse Document Erequency)(Sierra, 2019). Banyak cara menentukan nilai yang tepat dari kedua statistic yang ada. Dalam kasus trem frequency tf (t, d), cara paling sederhana adalah dengan menggunakan raw frequency di dalam dokumen, yaitu beberapa kali trem t muncul di dokumen d. jika menyatakan raw frequency t sebagai f (t, d), maka sekama tf yang sederhana adalah tf (t,d) = f (t,d).

Rumus *Term Frequency* pada persamaan 2.3

$$TF_t = f \qquad (2.3)$$

Keterangan:

TF : Frekuensi kemunculan kata dalam satu dokumen.

F : Jumlah kata pada satu dokumen.

Rumus *Invers Document Frequency* (IDF) terdapat pada persamaan 2.4 sebagai berikut:

$$IDF = Log\ \frac{N}{DF} \qquad (2.4)$$

Keterangan:

N : Jumlah Dokumen

DF : Nilai TF

Rumus TF-IDF terdapat pada persamaan 2.5

$$TF.IDF = TF \ x \ IDF \quad (2.5)$$

Keterangan:

TF : Nilai TF

IDF : Nilai IDF

8. *Lexicon Based*

Lexicon based merupakan kamus atau leksion yang digubakan untuk pemilihan kata pada data atau dokumen. Dalam implementasinya, tersedia dua kamus yaitu kamus dengan Kumpulan kata yang bersentimen positif dan kamus dengan Kumpulan kata yang bersentimen negatif yang digunakan untuk menjadi *wordlist*(Alvianda & Pandu Adikara, 2019). Metode analisis dari metode *lexion based* adalah VADER (*Valance Dictonary and Sentiment Reasoner)*. Vader digunakan untuk menganalisis data berdasarkan *lexicon* (kamus). Hasil dari vader berupa kelas polaritas positif, netral, netral dan negatif dengan tambahan *compound score* atau skor total. *Vader Sentimen* yang terkait dengan sinonim dan akronim serta kata berbahasa inggris(Rachmadana Ismail et al., 2023).

Leksikal merupakan kamus yang digunakan bahasa pokok dalam metode *lexion based.* Untuk mendeteksi klasifikasi atau sentiment, pada penelitian ini memanfatkan *libarary python* dengan *score polarty* < 0 adalah

sentiment negatif, *score polarity* = 0 adalah sentiment netral, dan *score polarity* > 0 adalah sentiment positif. Untuk proses klasifikasi sentiment dapat dilakukan dengan persamaan 2.6 berikut:

$$S_{positive} \sum_{iet}^{n} positive\ score_i \qquad (2.6)$$

$$S_{positive} \sum_{iet}^{n} negative\ score_i \qquad (2.7)$$

Diamana $S_{positeve}$ adalah bobot dari kalimat yang didapatkan melalui penjumlahan n skor polaritas kata opini positif dan $S_{negative}$ adalah bobot dari kalimat yang didapatkan melalui penjumlahan n skor polaritas kata opini negatif. Bobot pada tiap kalimat ini yang akan digunakan sebagai acuan untuk melakukan proses perbandingan. Sehingga dalam satu kalimat akan diketahui total jumlah nilai positif $S_{positif}$ dan juga nilai negatif $S_{negative}$ dari tiap-tiap kata penyusunannya. Dari persamaan nilai sentiment dalam satu kalimat maka di peroleh persamaan 3 untuk menentukan orientasi sentiment dengan perbandingan jumlah nilai positif, negatif dan netral.

9. Evaluasi

Tahap evaluasi yang dikerjakan dengan memakai teknik Confusion Matrix. Confusion Matrix adalah sebuah matriks yang menunjukan bagaimana sistem klasifikasi berbasis data bekerja(Sujatmiko & Seniwati, n.d.). Pada Langkah ini, perhitungan dilakukan guna menentukan nilai akurasi, presisi dan recall. Berikut adalah algoritma yang digunakan untuk mengukur akurasi Tabel 2.4 menunjukan ukuran evalusi model klasifikasi.

Gambar 2. 2 Ukuran Evaluasi Model Klasifikasi

		Predicted values	
Actual		1 (Positive)	0 (Negative)
Value	1 (Positive)	TP (True Positive)	FN (false Negative)
	0 (Negative)	FP (False Positive)	TN (True Negative)

Keterangan :

a. *True Positives* (TP) : Kelas kata terprediksi benar bernilai positif
b. *True Negatives* (TN) : Kelas kata negatif terprediksi negatif
c. *False Positives* (FP) : Kelas negatif terprediksi positif
d. *False Negatives* (FN) : Kelas positif terprediksi negatif

Dalam tahap evaluasi perhitungan akan diuji dengan akurasi, *presisi*, *recall* dan *f1-score* yang dijelaskan sebagai berikut.:

1) Akurasi : ukuran yang menunjukan sejauh mana hasil suatu pengukuran, prediksi, atau klasifikasi sesuai dengan nilai atau keadaan yang sebenarnya. Dalam konteks evaluasi kinerja, akurasi didefenisikan sebagau proporsi antara jumlah hasil yang benar dengan total keseluruhan pengujian atau observasi. Akurasi dapat di hitung dengan persamaan:

$$Akurasi = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (2.8)$$

2) *Precision* (Presisi): *Precision* mengukur seberapa banyak prediksi yang positif benar-benar positif dibandingkan dengan jumlah prediksi positif yang dilakukan oleh model. Dalam konteks analisis sentimen, precision menunjukkan seberapa banyak dari semua teks yang diklasifikasikan sebagai positif yang benar-benar memiliki sentimen positif. Presisi dapat dihitung dengan persamaan :

$$Presisi = \frac{TP}{(TP+FP)} \quad (2.9)$$

3) *Recall* (*Recall*): *Recall* mengukur seberapa banyak dari semua data yang sebenarnya positif berhasil diidentifikasi oleh model sebagai positif. *Recall* menunjukkan kemampuan model untuk menangkap semua contoh positif dari data. Dalam analisis sentimen, recall membantu dalam menilai seberapa baik model dapat menangani kasus-kasus sentimen positif yang ada dalam dataset. *Recall* dapat dihitung dengan persamaan :

$$Recall = \frac{TP}{(TP+FN)} \quad (3.0)$$

4) *F1-Score: F1-score* adalah metrik yang menggabungkan *precision* dan *recall* ke dalam satu nilai tunggal. *F1-score* adalah rata-rata harmonis dari precision dan *recall*, dan memberikan gambaran yang lebih seimbang tentang performa model, terutama ketika ada ketidakseimbangan kelas. *F1-score* sangat berguna ketika kita membutuhkan keseimbangan antara *precision* dan *recall* dan menghindari *trade-off* antara keduanya. *F1-score* dapat dihitung dengan persamaan :

$$FI - score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (3.1)$$

Atau

$$FI - score = 2 * \frac{precision * recall}{precision + recall} \quad (3.2)$$

BAB III
METODE PENELITIAN

A. Jenis Penelitian

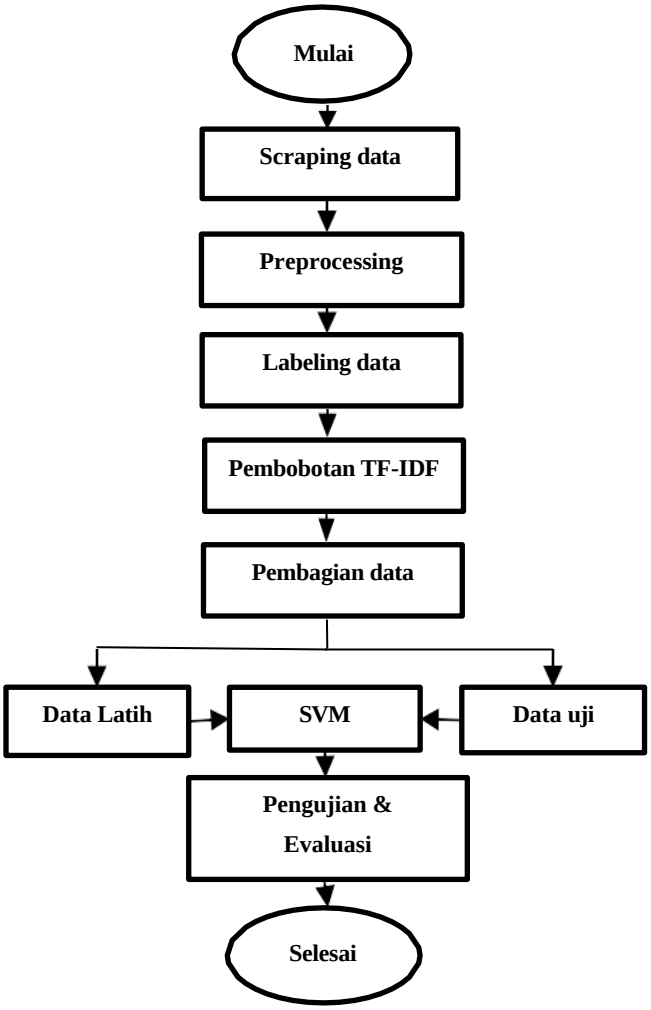
Jenis penelitian yang digunakan dalam penelitian ini adalah kuantitatif. Dalam penelitian ini proses pengambilan data dilakukan dengan menggunakan program *crawling* data tang menghasilkan data sebanyak 1.398 data. Program *crawling* data menggunakan *text editor* jupyter notebook dan bahasa pemograman *pyhton* dengan memanfatkan modul yang tersedia di *python* taitu *snscrape.* Dalam penelitian ini, metode terapan diterapkan untuk menganalisis sentiment terhadap kebijakan kenaikan PPN 12% berdasarkan respon pengguna media sosial X dengan menggunakan Teknik *Support Vector Machine* (SVM).

B. Sumber Data

Data dalam penelitian ini bersumber dari media sosial X. Proses pengambilan data dilakukan dengan teknik *crawling* menggunakan *Google Colab* dan bahasa pemrograman *Python*. Rentang waktu pengambilan data dimulai dari desember 2024. Data yang dikumpulkan meliputi respons pengguna terhadap kebijakan kenaikan pajak pertambahan nilai 12%, dengan menggunakan kata kunci "kenaikan PPN 12%," dan "Masyarakat tolak kenaikan PPN 12%,”. Hanya postingan yang berbahasa Indonesia yang diambil untuk memastikan relevansi dalam konteks lokal.

C. Langkah-Langkah Penelitian

Penelitian ini dilakukan melalui beberapa tahapan utama yang meliputi *crawling* data dari media sosial X, *preprocessing* data, pelabelan menggunakan kamus *lexicon*, visualisasi data, pembobotan TF-IDF, klasifikasi *Support Vector Machine* (SVM) dan evaluasi.



Gambar 3.1 Kerangka Pemikiran

D. Uraian Penelitian

1. Scraping data

Scraping data data merupakan cara yang digunakan secara otomatis guna mengumpulkan data yang terstruktur dari adanya web penggunaan aplikasi maupun kode program tertentu. Metode ini digunakan guna mengumpulkan setiap dataset dari media sosial twitter (X) yang dapat menghasilkan suatu data yang telah tersusun data yang dikumpulkan menggunakan scaping web bersama dengan Google Colaboratory selanjutnya dapat di unduh sebagai file CSV , yang telah disesuaikan dengan jumlah komentar dari pengunduh yang di perlukan.

2. Preprocessing

Preprocessing merupakan satu tahap awal yang dilakukan guna mempersiapkan data yang akan di proses lebih lanjut. Dalam tahap ini di kumpulkan data melalui beragam proses pengolahan teks, seperti membersihkan huruf kapital, teks, mengeluarkan sopword serta menyesuaikan. Tahap preprocessing bertujuan untuk Menyusun dataset secara terstruktur dan bersih sehingga mempermudah proses analisis. Tahap preprocessing meliputi:

a. Cleaning

Tahap ini dilakukan menghilangkan semua atribut di luar huruf alfabet dengan tujuan mengurangi karakter atau symbol yang tidak penting guna dilakukan analisis.

b. Case folding

Tahap berikutnya merupakan pengubahan seluruh huruf kapital dalam teks menjadi huruf kecil guna menyeragamkan format teks

c. Tokenizing

Tahap ini guna membagi kalimat dalam datset membentuk kata-kata kecil guna lebih mudah di proses.

d. Stop removal

Langkah ini menyaring kata kata atau istilah yang tidak memiliki makna penting dalam dokumen teks untuk meningkatkan efesiensi analisis.

e. Stemming

Proses ini dilakukan untuk menghapus imbuhan kata guna mengembalikan kata ke bentuk aslinya.

3. Labeling data

Dataset yang dikumpulkan dengan memanfaatkan metode scraping web selanjutnya diproses dalam tahap pelabelan. Pada langkah tersebut, setiap komentar yang ditemukan dalam kumpulan data dilakukan evaluasi serta diberikan label guna membagi data yang terbagi menjadi dua, yakni negatif dan positif(Alhaq et al., n.d.-c). Tujuan dari pelabelan ini guna menentukan kelas dari setiap komentar yang ada dalam dokumen, apakah mereka

mengandung sentimen positif dan negatif. Sementara, pada kategori negatif dapat menunjukkan hasil ketidakpuasan, ketidaksetujuan, serta ketidaksenangan masyarakat. Dalam kategori positif, dapat menunjukkan hasil penghargaan kepuasan, kesenangan, serta kebahagian masyarakat

Setelah data melalui tahap *preprocessing*, langkah berikutnya adalah pelabelan sentimen, di mana setiap entitas teks atau dokumen diberikan label yang menggambarkan polaritas sentimen apakah positif atau negatif dengan menggunakan *lexicon based*. (Hamka et al., 2022). Jika kata tersebut memiliki nilai sentimen yang relevan, nilai tersebut akan digunakan untuk menentukan label sentimen keseluruhan dari teks tersebut.

Dengan menggunakan metode ini, polaritas sentimen (positif atau negatif) dapat diidentifikasi secara lebih sistematis dan konsisten, yang selanjutnya memungkinkan analisis yang lebih mendalam dan akurat terhadap data yang dianalisis(Sanhaji et al., 2024). *Polarity score* dapat dilihat pada Tabel 3.7.

Tabel 3. 1 *Polarity Score* Pada *Lexicion Based*

Senitmen	*Polarity Score*
Positive	> 0
Negative	< 0

Berdasarkan data pada tabel 3.7 merupakan polarity score dengan kriteria bahwa sebuah tweet dianggap memiliki sentimen positif jika nilai

polatiry lebih dari sama dengan nol (> 0), jika nilai polarity kurang dari nol (< 0) tweet dianggap memiliki sentimen negatif.

4. Pembobotan TF-IDF

Pembobotan TF-IF adalah metode yang digunakan untuk menghitung bobot setiap kata yang sering muncul dalam proses penelurusan informasi (information retrieval). Metode ini dikenal efesien, sederhana, dan mampu memberikan hasil yang akurat. Prinsip kerja TF-IDF adalah dengan memberikan bobot pada pada hubungan suatu kata (trem) terhadap dokumen tertentu. TF-IDF merupakan ukuran statistic yang berguna untuk menilai seberapa penting sebuah kata dalam suatu dokumen atau Kumpulan dokumen. Dalam konteks dokumen tanggal, setiap kalimat dapat dianggap sebagai sebuah dokumen. Frekuensi kemunculan kata dalam suatu dokumen menunjukan Tingkat kepentingan kata tersebut dalam dokumen yang bersangkutan. Disisi lain, jumlah dokumen yang mengandung kata tersebut mencerminkan seberapa umum kata itu digunakan. Semakin sering sebuah kata muncul dalam dokumen tertentu, semakin tinggi bobotnya. Namun jika kata tersebut muncul di banyak dokumen, bobotnya justru semakin rendah.

5. Pembagian data

Pembagian data uji adalah Langkah penting dalam proses Pembangunan *machine learning*, termasuk pada analisis sentiment menggunakan algoritma *support vector machine* (SVM). Data di bagi menjadi dua yaitu :

a. Data latih dataset yang digunakan untuk melatih model. Model belajar dari polah dalam data ini untuk memahami hubungan antara input (fitur) dan output (label sentiment : positif, negatif, atau netral).
b. Data uji bagian dari dataset yang digunakan untuk menguji performa model setelah proses pelatihan disebut data pengujian. Data ini tidak dilibatkan selam tahao pelatihan, sehingga model belum pernah mengenalnya sebelumnya. Hal ini bertujuan untuk mengevaluasi seberapa baik model dapat menggeneralisasi pada data baru.

6. *Support vector machine* (SVM)

Support vector machine (SVM) adalah sebuah metode pembelajaran mesin yang digunakan untuk masalah dan regresi. SVM bekerja dengan membangun sebuah hyperlane atau garis pemisah optimal yang menghasilkan jarak antara dua kelas data yang berbeda. Hyperlane tersebut digunakan untuk mengklasifikasikan data baru berdasarkan posisinya terhadap hayperlane tersebut(Alhaq et al., n.d.-c). Keungulan utama SVM adalah kemampuannya dalam menangani data yang kompleks dan memiliki dimensi tinggi. SVM dapat mengatasi masalah overfiting dengan menggunakan fungsi kernel yang memetakan data ke dalam ruangan fitur yang lebih tingggi, di mana data menjadi lebih terpisah secara linear. Beberapa jenis fungsi kernel yang umum digunakan adalah kernel yang umum digunakan adalah kernel linear, kernel polynominal, dan kernel Gaussian.

7. Pengujian dan Evaluasi

Setelah tahap klasifikasi menggunakan Support Vector Machine (SVM) selesai, langkah selanjutnya adalah evaluasi hasil klasifikasi untuk menilai kinerja model. Evaluasi dilakukan menggunakan *confusion matrix* pada tabel 2.4. Pengujian dilakukan untuk mengukur seberapa efektif model dalam mengklasifikasikan data ke dalam kategori yang benar. Hasil dari *confusion matrix* divisualisasikan dalam bentuk *heatmap* yang menunjukkan jumlah setiap kategori dalam matrix diantaranya, akurasi, presisi, recall dan F1-Score. Perhitungan dilakukan dengan berdasarkan persamaan 2.8, 2.9, 3.0, 3.1.

DAFTAR PUSTAKA

Alhaq, Z., Mustopa, A., & Santoso, J. D. (n.d.-a). *PENERAPAN METODE SUPPORT VECTOR MACHINE UNTUK ANALISIS SENTIMEN PENGGUNA TWITTER*.

Alhaq, Z., Mustopa, A., & Santoso, J. D. (n.d.-b). *PENERAPAN METODE SUPPORT VECTOR MACHINE UNTUK ANALISIS SENTIMEN PENGGUNA TWITTER*.

Alhaq, Z., Mustopa, A., & Santoso, J. D. (n.d.-c). *PENERAPAN METODE SUPPORT VECTOR MACHINE UNTUK ANALISIS SENTIMEN PENGGUNA TWITTER*.

Alvianda, F., & Pandu Adikara, P. (2019). *Analisis Sentimen Konten Radikal Di Media Sosial Twitter Menggunakan Metode Support Vector Machine (SVM)* (Vol. 3, Issue 1). http://j-ptiik.ub.ac.id

buku python. (n.d.).

Firdaus, H., & Budiman, Y. N. (2024, November 26). *Tolak PPN 12% Viral di X, Apakah Seruan Praktik Frugal Living Efektif?* https://yoursay.suara.com/kolom/2024/11/26/092032/tolak-ppn-12-viral-di-x-apakah-seruan-praktik-frugal-living-efektif

Hakim, B. (2021). Analisa Sentimen Data Text Preprocessing Pada Data Mining Dengan Menggunakan Machine Learning. *JBASE - Journal of Business and Audit Information Systems*, *4*(2). https://doi.org/10.30813/jbase.v4i2.3000

Hamka, M., Alfatari, N., & Ratna Sari, D. (2022). Analisis Sentimen Produk Kecantikan Jenis Serum Menggunakan Algoritma Naïve Bayes Classifier. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *4*(1), 64. https://doi.org/10.30865/json.v4i1.4740

Husada, H. C., & Paramita, A. S. (2021). Analisis Sentimen Pada Maskapai Penerbangan di Platform Twitter Menggunakan Algoritma Support Vector Machine (SVM). *Teknika*, *10*(1), 18–26. https://doi.org/10.34148/teknika.v10i1.311

Kristovani Siagian, J. (2024a). *ANALISIS SENTIMEN MASYARAKAT INDONESIA TERHADAP RENCANA KENAIKAN PPN MENJADI 12% DI MEDIA SOSIAL X DENGAN METODE NAÏVE BAYES*. *3*(2).

Kristovani Siagian, J. (2024b). *ANALISIS SENTIMEN MASYARAKAT INDONESIA TERHADAP RENCANA KENAIKAN PPN MENJADI 12% DI MEDIA SOSIAL X DENGAN METODE NAÏVE BAYES*. *3*(2).