ISSN: 2339-2541
JURNAL GAUSSIAN, Volume X, Nomor X, Tahun XXXX, Halaman XX - XX
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

NATURAL LANGUAGE PROCESSING UNTUK SENTIMEN ANALISIS TERHADAP INSTANSI BEA CUKAI MENGGUNAKAN LONG SHORT-TERM MEMORY

La Ode Muhammad Yudhy Prayitno[1*], Diki Ardika Wiratama[2], Putri Rahayu[3]
[1,2,3] Kendari
***e-mail** : mr.yudhyt@gmail.com, ardykaaw26@gmail.com, asahisarin@gmail.com

DOI: 10.14710/J.GAUSS.XX.X.XX-XX

Article Info:
Received:
Accepted:
Available Online:

Keywords:
Bea Cukai, Sentiment; Natural Language Processing; Long Short-Term Memory; Sentiment Analysis

Abstract: Customs is a government agency responsible for supervising and controlling the flow of goods in and out of the country. Recently, Customs has come under public scrutiny due to controversial policies. This study uses Natural Language Processing (NLP) to analyze sentiment towards Customs based on CNBC Indonesia news articles from 2019 to 2024, totaling 1,783 articles. Data was processed through tokenizing, stopwords removal, stemming, and sentiment scoring using the Indonesian Sentiment Lexicon (InSet). A Long Short-Term Memory (LSTM) model was used for sentiment analysis, achieving a highest accuracy of 71.43%. Word cloud analysis highlights news focus on economic contributions, trade issues, and legal problems. This study helps understand public perceptions of Customs in the context of economics and regulation.

1. PENDAHULUAN

Pemerintah adalah entitas organisasi yang bertanggung jawab atas sistem pemerintahan dalam mengatur dan memimpin suatu negara atau wilayah, serta memiliki peranan penting dalam menyusun kebijakan, melaksanakan undang-undang, dan melayani warganya. Salah satu fungsi utama pemerintah adalah menyebarkan informasi dan mengomunikasikan tujuan, rencana kerja, serta kebijakan kepada publik melalui media tradisional, konvensional, dan baru (PERMENPANRB, 2012). Penggunaan teknologi internet atau media baru memungkinkan komunikasi langsung dan cepat dengan semua pihak, yang saat ini sangat disukai oleh masyarakat.

Perkembangan internet di Indonesia menunjukkan kemajuan besar, seperti yang ditunjukkan oleh data survei yang dirilis oleh Asosiasi Penyelenggara Jasa Internet Indonesia (2023), di mana jumlah pengguna internet di Indonesia mencapai 215,63 juta pada periode 2022–2023, meningkat sebesar 2,67% dari tahun sebelumnya. Kehidupan masyarakat modern telah banyak berubah karena kemajuan pesat dalam bidang sains dan teknologi, membuat masyarakat semakin kritis terhadap kondisi saat ini dan menuntut pemerintah untuk memenuhi kebutuhan mereka dalam segala aspek. Untuk membentuk pemerintahan yang baik dan transparan, diperlukan proses kerja sama dan partisipasi, termasuk umpan balik dari masyarakat yang mencakup kritik dan saran.

Media sosial tidak hanya menyediakan platform bagi individu untuk menyebarkan informasi, tetapi juga sebagai alat untuk mendapatkan umpan balik dari masyarakat. Menurut laporan We Are Social (2023), 167 juta orang di Indonesia aktif menggunakan media sosial pada Januari 2023, atau 60,4% dari total penduduk. Dengan akses yang mudah dan murah, masyarakat cenderung memanfaatkan media sosial untuk menyampaikan keluhan, saran, dan mencari informasi.

Berita adalah informasi terkini yang menarik minat banyak orang dan dapat diakses melalui koran, majalah, televisi, radio, serta portal berita online. Dampak berita terhadap

pembaca bisa positif, negatif, atau netral. Oleh karena itu, diperlukan sistem yang bisa mengategorikan sentimen berita untuk membantu pembaca dalam menilai berita secara bijaksana. Proses menyeleksi dan menganalisis berita secara manual memakan waktu dan tenaga, terutama dengan meningkatnya jumlah artikel yang dimuat di berbagai situs berita.

Penelitian ini menerapkan algoritma Long Short-Term Memory (LSTM), sebuah varian dari Recurrent Neural Network (RNN) yang telah dimodifikasi dengan menambahkan sel memori untuk menyimpan informasi jangka panjang (Souma, Vodenska, dan Aoyama, 2019). LSTM memiliki kelebihan dibandingkan RNN karena mampu mengelola error dan mengingat serta melupakan output yang akan diproses kembali menjadi input (Zhang, 2016). Fokus penelitian ini adalah pada sentimen positif dan negatif dalam berita tentang instansi Bea Cukai, dengan data diambil dari artikel berita CNBC Indonesia melalui teknik web scraping. Pendekatan ini cocok digunakan dalam analisis berita, seperti yang diterapkan dalam perdagangan saham berdasarkan sentimen berita (Tetlock, 2007).

Penelitian ini bertujuan untuk mengembangkan sistem klasifikasi sentimen otomatis yang dapat membantu mengidentifikasi persepsi publik terhadap instansi Bea Cukai, serta memberikan wawasan yang bermanfaat bagi pihak terkait dalam memahami dan merespons opini publik secara efektif.

2. TINJAUAN PUSTAKA

Bea Cukai adalah instansi pemerintah yang bertanggung jawab atas pengawasan dan pengendalian lalu lintas barang yang masuk dan keluar dari suatu negara. Instansi ini memainkan peran krusial dalam mengamankan perbatasan negara, melindungi industri dalam negeri, serta memungut bea masuk dan pajak impor. Tugas utama Bea Cukai meliputi pengawasan terhadap barang-barang yang diimpor dan diekspor, penegakan hukum terhadap penyelundupan, serta penerapan kebijakan perdagangan internasional. Melalui berbagai kebijakan dan regulasi, Bea Cukai berupaya untuk menciptakan lingkungan perdagangan yang aman, efisien, dan adil bagi semua pihak yang terlibat.

Namun, dalam beberapa waktu terakhir, Bea Cukai menjadi sorotan publik akibat berbagai kebijakan kontroversial yang mereka terapkan. Sentimen masyarakat terhadap instansi ini seringkali dipengaruhi oleh pemberitaan media, baik yang bersifat positif maupun negatif. Misalnya, kebijakan baru mengenai batasan impor barang tertentu dapat memicu reaksi keras dari masyarakat yang merasa dirugikan. Di sisi lain, upaya Bea Cukai dalam mengungkap kasus penyelundupan besar-besaran dapat memperoleh apresiasi dan dukungan dari publik. Analisis sentimen terhadap berita-berita ini penting untuk memahami persepsi masyarakat terhadap Bea Cukai dan untuk membantu instansi tersebut dalam merespons umpan balik publik secara lebih efektif.

Natural Language Processing (NLP) adalah cabang dari kecerdasan buatan yang berfokus pada interaksi antara komputer dan manusia melalui bahasa alami. Teknik NLP digunakan untuk menganalisis, memahami, dan menghasilkan bahasa manusia dengan cara yang bermanfaat. Salah satu aplikasi utama dari NLP adalah analisis sentimen, yaitu proses mengidentifikasi dan mengkategorikan opini yang diekspresikan dalam suatu teks menjadi sentimen positif, negatif, atau netral (Liu, 2012).

Analisis sentimen termasuk kategori didalam text mining yang menggali data didalam suatu sumber informasi bisa diweb misal didalam penelitian ini adalah CNBC Indonesia dan dalam permasalahan penilitan ini adalah analisis sentimen sebuah metode untuk

menganalisis opini publik terhadap suatu objek dalam implementasinya banyak sekali cara misalnya disini analisis sentimen.

Proses mengolah data tekstual secara otomatis untuk mengidentifikasi dan mengkategorikan kalimat opini juga dikenal sebagai analisis sentimen atau opinion mining. Hasil dari analisis sentimen digunakan untuk mengidentifikasi opini positif dan atau negatif yang tersirat dalam teks (Liu, 2015). Text mining mengacu pada proses mengekstraksi informasi dari sumber data dengan mengidentifikasi dan mengeksplorasi pola yang menarik seperti klasifikasi teks, ekstraksi informasi, dan ekstraksi kata (Feldman dan Sanger, 2007). Solusi untuk masalah seperti memproses, menyortir, mengkategorikan, dan menganalisis data besar yang tidak terstruktur dapat ditemukan melalui text mining (Nurhuda dan Sihwi, 2014).

Transformasi data tekstual menjadi format yang lebih sederhana agar mudah dibaca oleh sistem sebagai persiapan untuk pemrosesan selanjutnya disebut sebagai text preprocessing (Indraloka dan Santosa, 2017). Tahap ini dilakukan agar pengolahan data dapat dibuat lebih efisien pada tahap awal text mining. Tahapan text pre-processing pada penelitian ini meliputi case folding, remove URL, unescape HTML, remove mention, remove punctuation, remove number, remove duplicate, dan normalisasi kata. Word tokenizing adalah proses memecah teks menjadi unit-unit kata yang lebih kecil, yang disebut token. Proses ini merupakan langkah awal dalam pemrosesan bahasa alami (NLP) karena memungkinkan analisis lebih lanjut pada tingkat kata. Tokenizing sangat penting dalam analisis sentimen berita karena membantu dalam memisahkan kata-kata sehingga setiap kata dapat dianalisis secara individual untuk menentukan sentimennya. Proses tokenizing juga membantu dalam mengidentifikasi pola kata dan frekuensi kata dalam korpus teks berita (Manning et al., 2008).

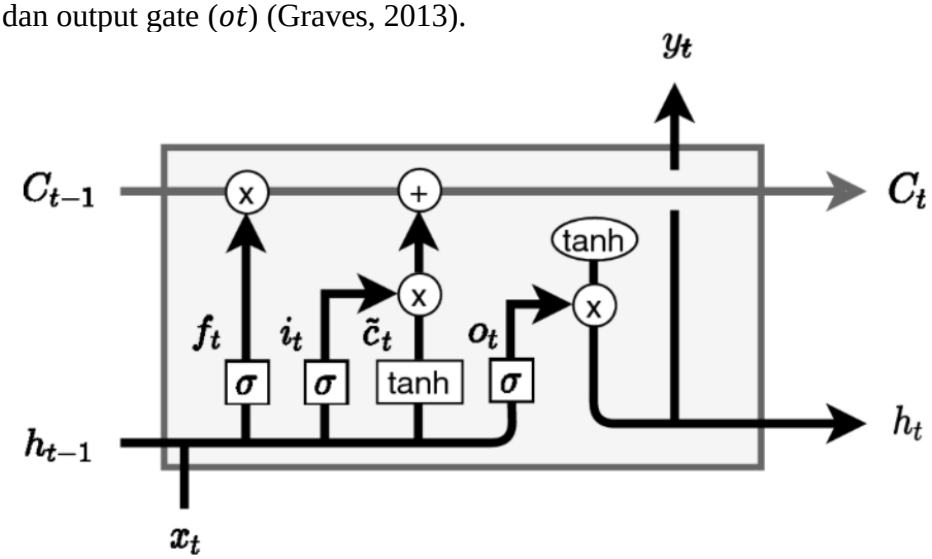
Stopwords removal berfungsi untuk menghilangkan kata-kata yang tidak memberikan kontribusi signifikan terhadap makna atau pesan dalam suatu teks. Kata-kata seperti "dan", "di", "yang", dan lain-lain biasanya tidak memiliki nilai informasi yang tinggi dan dapat diabaikan dalam analisis. Penggunaan stopwords removal terbukti dapat meningkatkan akurasi sistem klasifikasi sentimen karena mengurangi noise dalam data (Ghag dan Shah, 2015). Dalam analisis sentimen berita, menghilangkan stopwords membantu dalam fokus pada kata-kata yang benar-benar membawa makna sentimen. Stemming adalah proses mengubah kata berafiks, imbuhan, dan sufiks menjadi bentuk dasar atau kata dasar. Tujuan stemming adalah untuk menyatukan varian kata yang memiliki arti yang sama menjadi satu bentuk yang konsisten. Dalam konteks analisis sentimen berita, stemming membantu mengurangi dimensi fitur dan memperbaiki akurasi dengan memastikan bahwa kata-kata yang memiliki makna serupa diidentifikasi sebagai satu entitas (Porter, 1980). Hal ini penting karena kata-kata yang berimbuhan dapat memiliki bentuk yang berbeda tetapi makna yang sama.

Sentiment scoring bertujuan untuk melabelkan suatu pernyataan sehingga dapat digolongkan menjadi sentimen positif atau negatif berdasarkan kamus sentimen. Dalam penelitian ini, kamus yang digunakan adalah Indonesian Sentiment Lexicon (InSet), yang berisi kata-kata yang telah diberikan bobot sentimen dari 1 hingga 5 untuk sentimen positif dan -1 hingga -5 untuk sentimen negatif (Nielsen, 2011). InSet disusun menggunakan kumpulan kata dari tweets Indonesia yang terdiri dari 3.609 kata positif dan 6.609 kata negatif. Kamus sentimen seperti InSet sangat penting dalam analisis sentimen berita karena

mampu memberikan label sentimen secara otomatis pada teks, sehingga memudahkan proses klasifikasi sentimen.

Teknik yang dapat mengubah kata-kata individual dalam teks berita menjadi sebuah nilai vektor berupa bilangan riil dikenal sebagai word embedding (Brownlee, 2020). Word embedding memetakan kata ke dalam ruang vektor yang diwakili oleh vektor, dan setiap kata dalam dokumen dipetakan ke dalam vektor tersebut. Hasil word embedding berupa lookup table berbentuk matriks dengan dictionary size dan embedding size. Dictionary size merupakan ukuran kosakata dalam teks, sedangkan embedding size merupakan ukuran ruang vektor tempat kata-kata disematkan. Dalam konteks analisis sentimen berita, word embedding memungkinkan model untuk memahami konteks kata-kata dalam artikel berita dengan lebih baik, sehingga meningkatkan akurasi dalam pengenalan sentimen.

Long Short-Term Memory (LSTM) adalah varian dari Recurrent Neural Network (RNN) yang dirancang untuk mengatasi masalah vanishing gradient yang sering terjadi pada RNN standar. LSTM menggunakan sel memori yang dapat mempertahankan informasi dalam jangka waktu yang lama, memungkinkan jaringan untuk mengingat informasi penting dari urutan input yang panjang (Hochreiter & Schmidhuber, 1997). Algoritma LSTM telah berhasil diterapkan dalam berbagai aplikasi NLP, termasuk analisis sentimen, terjemahan bahasa, dan pengenalan suara. LSTM memiliki tiga jenis gates yaitu forget gate (ft), input gate (it), dan output gate (ot) (Graves, 2013).



Gambar 1 Arsitektur LSTM

Langkah pertama pada LSTM adalah memilih informasi apa yang akan dibuang dari cell state,hal ini dibuat oleh sigmoid layer yang disebut forget gate layer. Pada gambar dapat dilihat sel LSTM akan memproses $ht-1$ dan xt sebagai input. Langkah kedua dari cell LSTM adalah menentukan informasi yang akan disimpan di cell state. Proses ini memiliki dua bagian, pertama lapisan sigmoid menentukan nilai yang akan diperbarui dari cell state, lalu bagian kedua lapisan tanh membuat vektor dari kandidat baru, lalu keduanya digabung untuk melakukan pembaruan pada cell state. Cell state berfungsi untuk membawa informasi dari sel dibelakang ke sel-sel LSTM selanjutnya, pada setiap timestep cell state akan diperbarui dengan menggunakan forget gate dan input gate untuk menentukan informasi yang akan

dibuang ataupun ditambahkan kedalam cell state. Output gate berguna untuk menentukan output dari cell state sekarang. Pertama, lapisan sigmoid menentukan bagian dari cell state yang menjadi output. Lalu, lapisan tanh akan mengubah nilai cell state menjadi antara -1 dan 1, kemudian nilai dari lapisan sigmoid dan lapisan tanh dikalikan (Olah, 2015). Berikut persamaan-persamaan yang digunakan pada sel LSTM:

$$ft = \sigma(Wfxt + Ufht{-}1 + bf) \quad (1)$$

$$it = \sigma(Wixt + Uiht{-}1 + bi) \quad (2)$$

$$\bar{C}\,t = tanh(Wcxt + Ucht{-}1) + bc \quad (3)$$

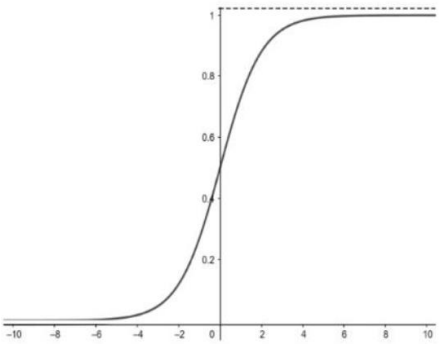
$$Ct = it \circ \bar{C}\,t + ft \circ Ct{-}1 \quad (4)$$

$$ot = \sigma(Woxt + Uoht{-}1 + bo) \quad (5)$$

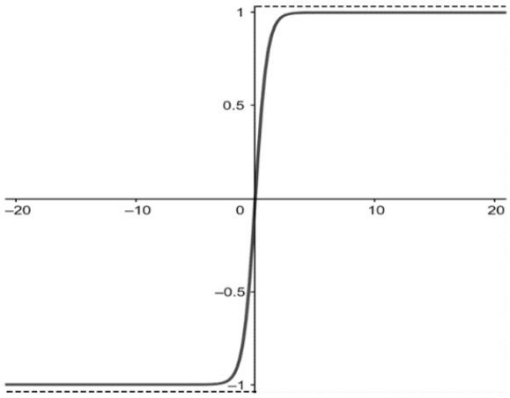
$$ht = ot \circ \tanh(Ct) \quad (6)$$

Simbol ft merupakan forget gate, it merupakan input gate, $\tilde{C}\ t$ merupakan cell state candidate. Ct merupakan cell state, ot merupakan output gate, ht merupakan hidden state, W merupakan matrik bobot, $ht{-}1$ merupakan hidden state sebelumnya, xt merupakan data input, b merupakan bias, σ merupakan fungsi aktivasi sigmoid, dan $tanh$ merupakan fungsi aktivasi tanh.

Sel LSTM menggunakan fungsi aktivasi sigmoid dan fungsi aktivasi tanh. Nilai input diubah ke dalam interval [0,1] oleh fungsi aktivasi sigmoid dan interval [-1,1] oleh fungsi aktivasi tanh. Gambar 2 dan 3 masing-masing menunjukkan grafik fungsi aktivasi sigmoid dan fungsi aktivasi tanh.



Gambar 2 Fungsi Aktivasi Sigmoid



Gambar 3 Fungsi Aktivasi Tanh

Persamaan 7 dan 8 adalah persamaan fungsi aktivasi sigmoid dan fungsi aktivasi tanh.

$$\sigma(x) = 1 / 1+e-x \text{ , } -\infty < x < \infty \quad (7)$$

$$tanh(x) = 2\ 1+e-2x - 1, -\infty < x < \infty \quad (8)$$

Loss function merupakan metode untuk mengevaluasi seberapa baik algoritma memodelkan suatu data (Li et al., 2019). Loss function memiliki kurva yang bertujuan memberi tahu cara mengubah parameter untuk membuat model lebih akurat. Metode cross entropy termasuk loss function untuk masalah klasifikasi yang memiliki output berupa nilai probabilitas antara 0 dan 1.

$$L = -y \log(\hat{y}) - (1 - y)\log (1 - (\hat{y})) \quad (9)$$

Simbol y merupakan nilai target, dan $\hat{y}$ merupakan nilai prediksi.

Optimasi Adam digunakan untuk mengoptimalkan fungsi tujuan dalam deep neural network. Dalam metode ini, proses perubahan parameter tergantung pada gradien, learning rate, nilai momen pertama dan kedua dari gradien. Berikut ini merupakan persamaan untuk optimasi Adam (Zhang, 2018):

$$gt = \nabla\theta f(\theta t-1) \quad (10)$$

$$mt = \beta 1 mt-1 + (1 - \beta 1) gt \quad (11)$$

$$vt = \beta 2 mt-1 + (1 - \beta 2) gt\ 2 \quad (12)$$

$$\hat{m}t = mt\ 1-\beta 1\ t \quad (13)$$

$$\hat{v}t = vt\ 1-\beta 2\ t \quad (14)$$

$$\theta t+1 = \theta - \eta\ \hat{m}t \surd\hat{v}t+\epsilon \quad (15)$$

Simbol θt merupakan parameter yang diperbaiki, η merupakan learning rate, $\hat{mt}$ merupakan bias dari estimator mt,$\hat{vt}$ merupakan bias dari estimator vt, dan ϵ merupakan epsilon.

Evaluasi kinerja klasifikasi bertujuan untuk mengetahui seberapa baik klasifikasi yang telah dibuat. Tingkat akurasi prediksi klasifikasi yang dihasilkan dievaluasi sebagai bagian dari proses pengukuran. Evaluasi kinerja klasifikasi pada penelitian ini dilakukan dengan menggunakan confusion matrix. Keberhasilan klasifikasi dalam mengenali tuple dari kelas yang berbeda dapat dievaluasi menggunakan confusion matrix (Han et al., 2012). Parameter yang diperoleh dari confusion matrix untuk menilai kinerja klasifikasi yaitu accuracy, precision, dan recall. Indikator pada confusion matrix dapat dilihat pada tabel 1.

Table 1 Confusion Matrix

		Kelas Prediksi	
		Positif	Negatif
Kelas Aktual	Postif	TP	FN
	Negatif	FP	TN

Word Cloud merupakan metode untuk menampilkan representasi visual dari data teks dengan ukuran yang berbeda. Dalam wordcloud, semakin besar ukuran kata menunjukkan semakin besar frekuensi kata muncul (Wardani et al., 2019). Visualisasi menggunakan word cloud bertujuan untuk membantu pengamat dalam melihat gagasan atau kata yang sering muncul dengan tampilan yang menarik.

3. METODE PENELITIAN

Data yang digunakan pada penelitian ini merupakan data kualitatif yang diperoleh dari proses scraping website CNBC Indonesia dengan keyword "Bea Cukai" yang diambil dari tahun 2019 hingga 2024. Konten artikel yang didapatkan berjumlah 2.611 artikel dan berbahasa Indonesia. Setelah dibersihkan dari duplikat, jumlah artikel yang digunakan menjadi 1.783 artikel. Variabel yang digunakan dalam pengambilan data meliputi headline, tanggal, link, dan konten artikel.

Langkah-langkah analisis yang dilakukan pada penelitian ini adalah sebagai berikut:

1. Screaping Berita CNBC
2. Explaratory Data Analysis
3. Preprocessing Data
4. Tokenizing
5. Sentiment Scoring
6. Stopwords removal
7. Stemming
8. Word embedding
9. Klasifikasi dengan algoritma Long-Short Term Memory.
    a. Membagi data menjadi data latih (training) dan data uji (testing)
    b. Membangun model LSTM

c. Mengevaluasi hasil kinerja klasifikasi menggunakan confusion matrix

10. Interpretasi dalam bentuk visual menggunakan wordcloud

4. HASIL DAN PEMBAHASAN

Data yang digunakan pada penelitian ini dikumpulkan dari website CNBC Indonesia dengan keyword "Bea Cukai" mulai dari tahun 2019 hingga 2024. Dari proses scraping tersebut, didapatkan 2.611 artikel berbahasa Indonesia. Setelah dilakukan proses penghilangan duplikat, jumlah artikel yang digunakan dalam penelitian ini adalah 1.783 artikel. Data artikel yang dikumpulkan disimpan dalam format CSV (Comma Separated Value) dengan variabel yang digunakan meliputi headline, tanggal, link, dan konten artikel.

Pre-processing data dilakukan dengan tujuan untuk mengubah data teks yang tidak terstruktur menjadi data yang terstruktur sehingga memudahkan tahap klasifikasi. Tahapan pre-processing yang dilakukan menggunakan teknik Natural Language Processing (NLP) adalah sebagai berikut:

1. Case Folding
2. Remove URL
3. Remove Punctuation dan Numbers
4. Remove Specific Phrases
5. Remove Duplicate

Dalam pembentukan sebuah NLP memerlukan sebuah pipeline dimana untuk model mesin yang dibuat dalam paper ini hanya menggunakan tahapan :

(1) Word tokenization
(2) Stopwords removal
(3) Stemming
(4) Split datasets
(5) Sentimen Scoring
(6) Word Embedding

Setelah proses pre-processing, langkah selanjutnya adalah tokenizing. Teks artikel dipecah menjadi unit-unit kata (token) untuk analisis lebih lanjut. Proses ini memungkinkan identifikasi pola kata dan frekuensi kata dalam korpus teks berita. Proses tokenizing dilakukan dengan memecah artikel berita menjadi unit-unit kata (token). Hal ini penting untuk memungkinkan analisis lebih lanjut terhadap pola kata dan frekuensi kata dalam korpus teks berita.

Setelah tokenizing, langkah berikutnya adalah menghilangkan stopwords. Stopwords seperti "dan", "di", "yang", dan kata-kata lain yang tidak memberikan kontribusi signifikan terhadap analisis sentimen dihapus. Langkah ini bertujuan untuk memfokuskan analisis pada kata-kata yang membawa makna sentimen yang lebih kuat dalam konteks artikel berita.

Hasil dari Stopwords Removal, artikel berita kemudian menjalani proses stemming. Stemming mengubah kata-kata menjadi bentuk dasarnya dengan menghapus afiks seperti imbuhan dan sufiks. Proses ini membantu dalam mengurangi variasi kata dan memastikan konsistensi dalam analisis sentimen.

Proses pelabelan data pada penelitian ini dilakukan dengan sentiment scoring menggunakan Indonesian Sentiment Lexicon (InSet). InSet menyediakan bobot sentimen untuk setiap kata, dengan rentang nilai antara 1 hingga 5 untuk sentimen positif dan -1 hingga -5 untuk sentimen negatif. Proses sentiment scoring ini memungkinkan klasifikasi otomatis

berdasarkan bobot sentimen yang telah ditentukan sebelumnya. Artikel dengan skor lebih besar dari 0 diberi label sentimen positif (1), sementara artikel dengan skor kurang dari 0 diberi label sentimen negatif (0).

Table 2 Hasil dari Tokenizing, Stopwords, Stemming, Sentiment Scoring

Content	Token	Stop_article	Stem_article	Prepos_article	polarity_score	polarity
pemerintah tengah bersiap merealisasikan ekspo...	['pemerintah', 'tengah', 'bersiap', 'merealisa...	['pemerintah', 'merealisasikan', 'ekspor', 'pa...	perintah realisasi ekspor pasir laut presiden ...	[perintah, realisasi, ekspor, pasir, laut, pre...	-12	negatif
……	……	……	……	……	……	……
……	……	……	……	……	……	……
direktur jenderal bea dan cukai kementerian ke...	['direktur', 'jenderal', 'bea', 'dan', 'cukai'...	['direktur', 'jenderal', 'bea', 'cukai', 'keme...	direktur jenderal bea cukai menteri uang askol...	[direktur, jenderal, bea, cukai, menteri, uang...	6	positif

Setelah proses sentiment scoring menggunakan Indonesian Sentiment Lexicon (InSet), langkah selanjutnya adalah mengubah teks artikel menjadi representasi numerik menggunakan teknik word embedding. Word embedding memetakan kata-kata ke dalam ruang vektor yang memungkinkan komputer untuk memahami hubungan semantik antara kata-kata tersebut. Teknik ini membantu meningkatkan kualitas analisis sentimen dengan menyandikan makna kata-kata dalam artikel berita ke dalam representasi matematis.

Mengubah kata-kata tersebut menjadi representasi numerik yang bisa diproses oleh model pembelajaran mesin. Fungsi Word Embedding dari Keras digunakan untuk mengubah teks menjadi urutan angka yang dapat dimasukkan ke dalam model LSTM. Hal ini penting karena model pembelajaran mesin bekerja dengan data numerik, bukan teks mentah. Kami menggunakan keras untuk mentokenisasi konten berita dengan memotong (memilih) 300 kata yang paling sering dipakai, dan menjadikan 300 kata tersebut (text corpus) menjadi vektor.

Lapisan pertama pada model LSTM yang akan dibangun adalah lapisan word embedding. Lapisan ini mengubah kata menjadi numerik seperti bobot nilai yang diinisialisasi secara acak menjadi lookup table. Lapisan word embedding terdapat beberapa parameter yaitu input_dim yang merupakan ukuran kosakata dalam teks yang memiliki ukuran 2.000 teks. Parameter lainnya yaitu output_ dim yang merupakan ukuran ruang vektor tempat kata-kata akan disematkan dengan ukuran 128 yang ditentukan dengan trial error. Parameter terakhir yaitu parameter input_length yang merupakan panjang dari urutan input yang berjumlah 300. Variabel input xt berupa 3D dengan ukuran (2000,300,128) menjadi output embedding berupa 2D dengan ukuran (300, 128) yang selanjutnya akan dimasukan ke lapisan LSTM.

Proses ini memungkinkan model LSTM untuk memahami konteks kata-kata dalam artikel secara lebih baik, dengan mempertimbangkan hubungan antar kata dalam kalimat.

Kemudian, dilakukan spatial dropout untuk mencegah overfitting dengan mengabaikan sebagian unit input selama pelatihan. Lalu, LSTM digunakan untuk memproses urutan kata-kata dan mengekstraksi fitur penting dari teks. Akhirnya, dense layer dengan satu unit digunakan untuk melakukan klasifikasi sentimen berdasarkan output yang dihasilkan oleh LSTM. Konfigurasi ini memungkinkan model untuk belajar dari data latih (training) dan menghasilkan prediksi sentimen yang akurat pada data uji (testing), yang dievaluasi menggunakan confusion matrix untuk mengukur kinerja klasifikasi secara keseluruhan. Interpretasi hasilnya disajikan dalam bentuk visual menggunakan wordcloud untuk memvisualisasikan kata-kata yang paling berpengaruh dalam menentukan sentimen positif dan negatif dalam artikel berita.

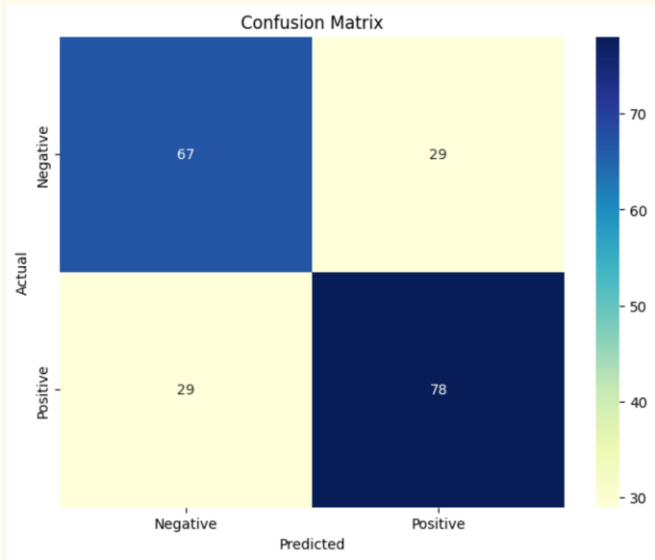
Layer (type)	Output Shape	Param #
embedding_18 (Embedding)	(None, 300, 128)	256,000
spatial_dropout1d_18 (SpatialDropout1D)	(None, 300, 128)	0
lstm_18 (LSTM)	(None, 100)	91,600
dense_18 (Dense)	(None, 1)	101

Gambar 4 Hasil Model LSTM

Data yang telah melalui proses pre-processing hingga word embedding dibagi menjadi data latih dan data uji dengan perbandingan 80% untuk data latih dan 20% untuk data uji. Pembangunan model LSTM dilakukan dengan eksplorasi hyperparameter menggunakan pendekatan trial and error. Pada penelitian ini, hyperparameter yang dieksplorasi meliputi jumlah unit LSTM, tingkat dropout, dan ukuran batch. Hyperparameter yang diuji mencakup jumlah unit LSTM sebesar 50, 100, dan 150, dengan tingkat dropout 0.2, 0.3, dan 0.4, serta ukuran batch 32 dan 64. Epoch untuk setiap model LSTM ditetapkan pada 100, namun penggunaan early stopping dilakukan untuk mencegah overfitting, yang dapat menghentikan pelatihan sebelum mencapai epoch maksimal jika terjadi penurunan signifikan dalam kinerja model. Proses propagasi maju pada LSTM dihubungkan dengan lapisan fully connected yang memiliki satu unit dengan fungsi aktivasi sigmoid. Optimizer yang digunakan adalah Adam, yang membantu dalam pembaruan bobot model berdasarkan metrik loss yang dihitung menggunakan binary cross entropy selama pelatihan data.

Dalam pengembangan model, hasil akurasi terbaik diperoleh dengan konfigurasi hyperparameter yang optimal, yaitu menggunakan 100 unit LSTM, tingkat dropout 0.2, dan ukuran batch 64. Model yang dihasilkan mencapai akurasi sebesar 71.43% pada data uji setelah pelatihan dengan 10 epoch. Hasil ini menunjukkan bahwa konfigurasi tersebut mampu menghasilkan model yang efektif dalam klasifikasi sentimen pada artikel berita setelah melalui proses pre-processing dan word embedding.

Model klasifikasi yang telah diuji dan dibangun menggunakan data latih akan dievaluasi kinerjanya. Evaluasi kinerja model pada penelitian ini menggunakan confusion matrix. Hasil dari confusion matrix untuk algoritma Long Short Term-Memory dapat dilihat pada gambar 5.



Gambar 5 Confusion Matrix

$$Accuracy = \frac{67+78}{67+29+29\_78} \times 100\% = 71.43\%$$

Hasil perhitungan overall accuracy menunjukkan tingkat akurasi model algoritma Long Short Term-Memory dalam mengklasifikasikan sentimen terkait instansi bea cukai pada website artikel berita adalah sebesar 71.43%.

Visualisasi data teks berupa teks berita dilakukan dengan menggunakan word cloud, visualisasi data dilakukan berdasarkan dua kelas sentimen yang telah ditentukan agar dapat diketahui kata-kata yang sering muncul pada masing-masing kelas sentimen. Word cloud untuk masing-masing kelas sentimen dapat dilihat pada Gambar 6 dan 7.

Gambar 6 Word Cloud Kelas Sentimen Positif

Gambar 7 Word Cloud Kelas Sentimen

Pada word cloud sentimen positif, kata-kata seperti "bea cukai", "rp", "triliun", "menteri", "uang", "sri", "mulyani", dan "indonesia" muncul dengan frekuensi yang tinggi. Hal ini menunjukkan bahwa berita atau artikel dengan sentimen positif sering kali menyoroti kontribusi bea cukai terhadap penerimaan negara, yang diukur dalam triliunan rupiah. Kata "menteri" dan "uang" yang sering muncul mengindikasikan peran penting Menteri Keuangan Sri Mulyani dalam mengelola dan memuji kinerja bea cukai. Kata "program" dan "dukungan" juga mencerminkan adanya inisiatif dan langkah-langkah positif yang diambil oleh pemerintah untuk mendukung dan memperbaiki sistem bea cukai di Indonesia.

Sebaliknya, pada word cloud sentimen negatif, kata "bea cukai", "batu bara", "menteri", "uang", "china", dan "rp triliun" menonjol. Kata "batu bara" yang dominan dalam word cloud negatif menunjukkan bahwa isu-isu seputar ekspor dan impor batu bara menjadi perhatian utama dalam konteks negatif. Hal ini mungkin terkait dengan kebijakan atau kontroversi mengenai perdagangan batu bara, termasuk dampak lingkungan atau ekonomi yang mungkin negatif. Kata "china" yang muncul juga menunjukkan adanya hubungan dengan negara tersebut dalam konteks yang kurang menguntungkan, mungkin terkait dengan perdagangan atau kebijakan bea cukai yang melibatkan China.

Selain itu, kata "tindak pidana" dan "pelanggaran" yang muncul di word cloud negatif menunjukkan bahwa berita dengan sentimen negatif sering kali terkait dengan isu-isu hukum dan pelanggaran yang melibatkan bea cukai, seperti penyelundupan atau korupsi. Kata-kata ini mencerminkan adanya masalah yang perlu diatasi untuk meningkatkan kinerja dan reputasi bea cukai di mata publik.

Secara keseluruhan, analisis word cloud menunjukkan bahwa sentimen positif terhadap bea cukai lebih banyak berfokus pada kontribusi ekonomi dan upaya peningkatan kinerja oleh pemerintah, sementara sentimen negatif lebih banyak menyoroti isu-isu kontroversial seperti perdagangan batu bara dan pelanggaran hukum. Kedua aspek ini penting untuk dipahami dalam rangka memperbaiki dan mengoptimalkan sistem bea cukai di Indonesia

5. KESIMPULAN

Dalam penelitian ini, kami berhasil menerapkan pendekatan NLP untuk menganalisis sentimen terhadap Bea Cukai berdasarkan data berita dari CNBC Indonesia. Langkah-langkah preprocessing seperti tokenizing, stopwords removal, dan stemming membantu mempersiapkan data untuk analisis sentimen. Sentiment scoring menggunakan InSet mengklasifikasikan berita menjadi sentimen positif dan negatif, sedangkan word embedding memberikan representasi vektor yang mewakili kata-kata dalam teks. Penggunaan model LSTM untuk klasifikasi sentimen menunjukkan hasil yang menggembirakan dengan akurasi tertinggi dicapai dengan hyperparameter yang dioptimalkan. Hasil perhitungan overall accuracy menunjukkan tingkat akurasi model algoritma Long Short-Term Memory dalam mengklasifikasikan sentimen terkait instansi Bea Cukai pada artikel berita dari CNBC Indonesia adalah sebesar 71.43%. Word cloud dari hasil analisis memberikan wawasan tambahan tentang fokus berita positif dan negatif terhadap Bea Cukai, yang penting untuk pemahaman dan pengambilan keputusan di bidang ini. Penelitian ini memberikan kontribusi dalam mengaplikasikan NLP dalam konteks analisis sentimen terhadap instansi pemerintah, khususnya terkait dengan kebijakan ekonomi dan perpajakan.

DAFTAR PUSTAKA

Brownlee, J. 2020. *Long Short-Term Memory Networks With Python Develop Sequence Prediction Models With Deep Learning*. Machine Learning Mastery.

Feldman, R., dan Sanger, J. 2007. *The Text Mining Handbook: Advanced approaches in Analyzing Unstructured Data*. New York: Cambridge University Press.

Ghag, K. V., dan Shah, K. 2015. *Comparative Analysis of Effect of Stopwords Removal on Sentiment Classification. International Conference on Computer, Communication and Control (IC4)*. India: Institute of Electrical and Electronics Engineers (IEEE).

Han, J., Kamber, M., dan Pei, J. 2012. *Data Mining: Concept and Techniques*. San Fransisco: Morgan Kaufmann Publishers.

Indraloka, D. S., dan Santosa, B. 2017. *Penerapan Text Mining untuk Melakukan Clustering Data Tweet Shopee Indonesia*. Jurnal Sains dan Seni ITS, 6(2): 6– 11. https://doi.org/10.12962/j23373520.v6i2.24419.

Kemenkes. 2021. *4 Manfaat Vaksin Covid-19 yang Wajib Diketahui*. https://upk.kemkes.go.id/new/4-manfaat-vaksin-covid-19-yang-wajib-diketahui.

Kemenkes. 2021. *Penjelasan WHO tentang Omicron, Varian Baru COVID-19*. https://covid19.go.id/p/berita/penjelasan-who-tentang-omicron-varian-baru-covid-19.

Li, C., Yuan, X., Lin, C., Guo, M., Wu, W., Yan, J., dan Ouyang, W. 2019. *AM-LFS: AutoML for loss function search*. Proceedings of the IEEE International Conference on Computer Vision, 2019-October(2), 8409–8418. https://doi.org/10.1109/ICCV.2019.00850.

Li, D., dan Qian, J., 2016. *Text Sentimen Analysis Based on Long Short-Term Memory*. Proceedings 1st IEEE International Conference on Computer Communication and the Internet. Wuhan, 13-15 Oktober, 471–475.

Li, S., dan Xu, J. 2018. A Recurrent Neural Network Language Model Based on Word Embedding. Springer, Cham, 368–377. https://doi.org/10.1007/978-3-030-01298-4_30.

Liu, B. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge: Cambridge University Press.

Manning, C., Raghavan, P., dan Schütze, H. 2009. *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Murthy, G. N., Allu, S. R., Andhavarapu, B., Bagadi, M. B. M. 2020. *Text based Sentiment Analysis using LSTM*. International Journal of Engineering and Technical Research V9(05). DOI: 10.17577/IJERTV9IS050290.

Nielsen, F. A. 2011. *A new ANEW: Evaluation of a word list for sentiment analysis in microblogs*. CEUR Workshop Proceedings, 718(March 2011), 93–98.

Nurhuda, F., dan Sihwi, S. W. 2014. *Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier*. ITSMART: Jurnal Ilmiah Teknologi Dan Informasi, 2: 35–42.

Olah, C., 2015. *Understanding LSTM Networks*. https://colah.github.io/posts/2015-08-Understanding-LSTMs.

Sembodo, J. E., Setiawan, E. B., dan Baizal, Z. A. 2016. *Data Crawling Otomatis pada Twitter*. Computational Science, School of Computing, Telkom University. October 2018, 11–16. https://doi.org/10.21108/indosc.2016.111.

Torres, J.F., Martínez-Álvarez, F. dan Troncoso, A. *A deep LSTM network for the Spanish electricity consumption forecasting*. Neural Comput dan Applic 34, 10533–10545 (2022). https://doi.org/10.1007/s00521-021-06773-2.

Wardani, F. K., Hananto, V. A., Nurcahyawati, V. 2019. *Analisis Sentimen Untuk Pemeringkatan Popularitas Situs Belanja Online di Indonesia Menggunakan Metode Naive Bayes (Studi Kasus Data Sekunder)*. JSIKA Vol. 08, No. 01.

Zhang, Z. 2018. *Improved Adam Optimizer for Deep Neural Networks*. IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), 2018, pp. 1-2, doi: 10.1109/IWQoS.2018.8624183.