

OPTIMIZING SENTIMENT ANALYSIS OF PRODUCT REVIEWS ON MARKETPLACE USING A COMBINATION OF PREPROCESSING TECHNIQUES, WORD2VEC, AND CONVOLUTIONAL NEURAL NETWORK

Fahry¹, Ema Utami², Sudarmawan³

^{1,2,3}Magister Teknik Informatika, Universitas Amikom Yogyakarta, Indonesia
Email: ¹fahry.1260@students.amikom.ac.id, ²ema.u@amikom.ac.id, ³sudarmawan@amikom.ac.id

(Naskah masuk: 09 Januari 2023, Revisi : 11 Januari 2023, diterbitkan: 10 Februari 2023)

Abstract

This research attempts to identify the most accurate and effective model in performing sentiment analysis on product reviews in marketplaces using preprocessing techniques, word2vec, and CNN. We collected 20,986 reviews from 720 products in a marketplace using scrap method, then cleaned and labeled the data to include 515 positive reviews, 490 negative reviews. We then performed preprocessing on the data using four different scenarios and identified word vector representation using word2vec. Subsequently, we applied the results of word2vec to the CNN architecture to classify sentiment in product reviews. After trying various variations of each technique, we found that a combination of the third preprocessing technique (case folding, punctuation removal, word normalization, and stemming), the second word2vec parameter combination (size 50, window 2, hs 0, and negative 10), and the fourth CNN parameter combination (kernel size 2, dropout 0.2, and learning rate 0.01) had the best accuracy of 99.00%, precision of 98.96%, and recall of 98.96%. We also found that the word normalization technique greatly helped to increase model accuracy by correcting improperly written or incorrect words in the reviews. Based on the evaluation of word2vec, the hs 0 method produced a higher average accuracy compared to the hs 1 method because the hs 0 method used negative sampling which helped the model understand the context of the trained words. In the CNN parameter, higher learning rates can cause the model to learn faster, but can also cause the model to be unstable, while lower learning rates can make the model more stable but can also cause the model's learning process to be slower.

Keywords: CNN, Preprocessing, Sentiment, Product Reviews, Word2vec.

OPTIMISASI ANALISIS SENTIMEN ULASAN PRODUK PADA MARKETPLACE DENGAN KOMBINASI TEKNIK PREPROCESSING, WORD2VEC, DAN CONVOLUTIONAL NEURAL NETWORK

Abstrak

Penelitian ini mencoba untuk mengidentifikasi model paling akurat dan efektif dalam melakukan analisis sentimen pada ulasan produk di *marketplace* menggunakan teknik *preprocessing*, word2vec, dan CNN. Kami mengumpulkan 20.986 ulasan dari 720 produk di *marketplace* shopee dengan metode *scrap*, kemudian membersihkan dan memberi label data sehingga tersisa 515 ulasan positif, 490 ulasan negatif. Kemudian, kami melakukan *preprocessing* dengan 4 skenario dan mengidentifikasi representasi vektor kata dengan word2vec. Selanjutnya, kami mengaplikasikan pada arsitektur CNN untuk melakukan klasifikasi sentimen ulasan produk. Setelah mencoba variasi dari masing-masing teknik, kami menemukan bahwa kombinasi teknik *preprocessing* ketiga (*case folding*, *punctuation removal*, *word normalizer*, dan *stemming*), kombinasi parameter word2vec kedua (*size 50*, *window 2*, *hs 0*, dan *negative 10*), dan kombinasi parameter CNN keempat (*kernel size 2*, *dropout 0.2*, dan *learning rate 0.01*) memiliki akurasi terbaik sebesar 99.00%, presisi 98.96%, dan *recall* 98.96%. Kami juga menemukan bahwa *word normalizer* membantu meningkatkan akurasi model karena memperbaiki kata yang salah dari ulasan. Berdasarkan hasil evaluasi word2vec, metode *hs 0* menghasilkan rata-rata akurasi yang lebih tinggi dibandingkan dengan metode *hs 1* karena metode *hs 0* menggunakan *negative sampling* yang membantu model memahami konteks kata yang dilatih. Pada parameter CNN, *learning rate* 0.001 dan 0.01 memiliki pengaruh lebih besar dibandingkan dengan *learning rate* 0.0001 yang memperlihatkan penurunan signifikan pada rata-rata akurasi model. *Learning rate* yang lebih tinggi dapat menyebabkan model belajar lebih cepat namun dapat menyebabkan model tidak stabil, sedangkan *learning rate* yang lebih rendah dapat menyebabkan model menjadi lebih stabil namun proses belajar model lebih lambat.

Kata kunci: CNN, Preprocessing, Sentimen, Ulasan Produk, Word2vec.

1. PENDAHULUAN

Analisis sentimen produk merupakan cara yang efektif untuk memahami bagaimana pelanggan menilai produk. Penelitian telah menunjukkan bahwa ulasan produk memberi pengaruh signifikan terhadap minat beli pelanggan [1] dan merupakan faktor terpenting yang mempengaruhi tingkat penjualan [2]. Ulasan produk di *marketplace* dapat memberikan informasi berguna bagi calon pelanggan untuk membuat keputusan belanja dan meminimalkan risiko membeli produk yang tidak sesuai atau tidak berkualitas [3], serta dapat menjadi masukan atau evaluasi bagi penjual [4]. Namun, analisis sentimen secara manual merupakan tugas yang sulit [5] karena banyaknya data yang harus dianalisis, ketidakkonsistenan ulasan pelanggan, keterbatasan manusia dalam melakukan analisis secara manual, dan potensi terpengaruh oleh bias individu. Salah satu solusi untuk masalah ini adalah dengan mengimplementasikan model klasifikasi *Convolutional Neural Network* (CNN) yang dapat memberikan akurasi tinggi dalam mengelompokkan sentimen pelanggan dalam bentuk ulasan produk [6].

CNN adalah salah satu model yang efisien menangkap maksud dari representasi teks ulasan dan mengklasifikasikan teks. Pada penelitian model CNN memberikan performa lebih baik dari model Naïve Bayes [7]. Penelitian lain membandingkan tiga model Naïve Bayes, SVM dan CNN dalam 3 topik data dan CNN memberikan hasil lebih baik terhadap ketiga topik data dibandingkan dengan SVM dan Naïve Bayes [8]. Meskipun CNN memiliki kemampuan yang luar biasa dalam klasifikasi teks, terdapat banyak cara untuk memperkuat performa dengan menggunakan *preprocessing* dan representasi teks yang optimal, seperti *word2vec*. Dengan mencoba berbagai kombinasi *preprocessing* NLP, parameter *word2vec*, dan parameter pada CNN, kita dapat mengetahui faktor-faktor yang paling penting untuk dipertimbangkan dalam klasifikasi teks. Hal ini juga dapat membantu menemukan representasi yang lebih baik dari teks, meningkatkan akurasi dan kemampuan CNN dalam mengklasifikasikan teks.

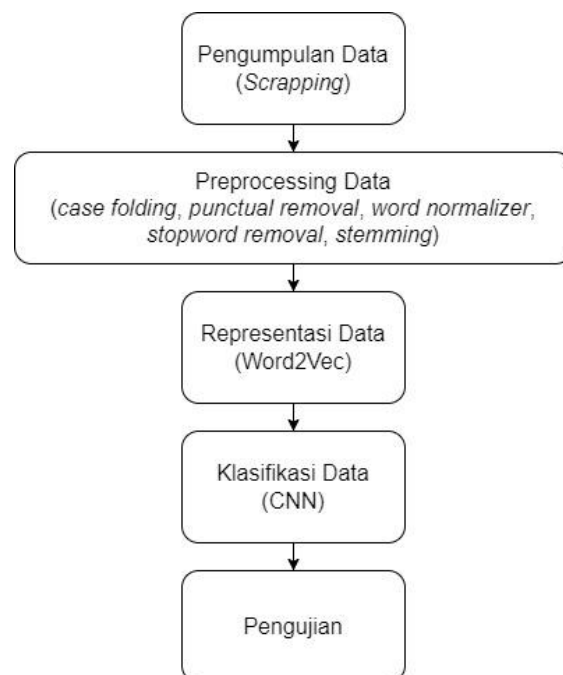
Pada penelitian [4] menyimpulkan penggunaan pendekatan NLP dalam pra-pemrosesan data meningkatkan akurasi klasifikasi menjadi 76,92%. Penelitian lain [3] menemukan bahwa klasifikasi ulasan Indonesia dengan perbaikan kata memiliki akurasi 87,60%. Penelitian [9] membandingkan kinerja word embedding Word2Vec, GloVe, dan FastText pada klasifikasi teks menggunakan CNN. Hasil menunjukkan bahwa ketiganya memiliki kinerja yang kompetitif dengan FastText yang terbaik, namun *word2vec* juga menunjukkan hasil yang baik. Penelitian sebelumnya menunjukkan penggunaan pendekatan NLP dan word embedding

seperti Word2Vec dapat meningkatkan akurasi klasifikasi sentimen menggunakan model CNN.

Penelitian ini dilakukan untuk mengetahui bagaimana analisis sentimen terhadap ulasan produk di *marketplace* menggunakan CNN. Ulasan tersebut kemudian akan diolah dengan menggunakan NLP dan direpresentasikan dengan *word2vec* sebelum dianalisis menggunakan CNN. Penelitian ini akan mengeksplorasi dan menguji berbagai kombinasi *preprocessing*, parameter pada *word2vec*, dan CNN untuk menemukan kombinasi terbaik yang dapat meningkatkan akurasi analisis sentimen.

2. METODE PENELITIAN

Untuk memberikan gambaran yang lebih lengkap tentang metode yang kami gunakan, pada bagian ini kami akan menjelaskan tahapan yang dilakukan mulai dari pengumpulan data, pengolahan data, hingga proses merepresentasikan data dengan *word2vec* sebagai input pada arsitektur CNN. Semua tahapan ini akan dijabarkan dibawah ini dan tahapan penelitian diilustrasikan pada gambar 1.



Gambar 1. Tahapan penelitian

2.1. Pengumpulan Data

Pengumpulan data ulasan produk dilakukan dengan metode scrap [10] dari *marketplace* shopee. Kemudian, data tersebut dibersihkan dari data duplikat dan kesalahan dalam pengambilan ulasan. Selanjutnya, data tersebut diberi label positif, negatif dan tidak relevan yang dilakukan oleh beberapa orang untuk mengurangi bias dan subjektivitas [11]. Proses selanjutnya, hanya data yang berlabel positif dan negatif yang akan digunakan.

2.2. Preprocessing Data

Preprocessing bertujuan membersihkan dataset dari *noise* untuk meningkatkan akurasi model [12]. Penelitian dengan NLP dapat meningkatkan akurasi sebesar 3% dibandingkan pada teks biasa [13]. Dalam tahap *preprocessing*, terdapat lima operasi fitur NLP yang dilakukan. Kami menggunakan empat skenario dengan variasi penerapan fitur berbeda untuk mengevaluasi pengaruh masing-masing fitur terhadap hasil analisis. Kami akan menganalisis hasil dari masing-masing skenario untuk menentukan yang paling efektif. Fitur NLP yang digunakan dalam adalah:

1. *Case folding* sebagai proses mengubah teks menjadi huruf kecil. Proses ini mengurangi dimensi data sehingga tidak membedakan kata dengan huruf besar dan huruf kecil.
2. *Punctuation removal* yaitu menghilangkan tanda baca yang memfokuskan penelitian hanya pada kata.
3. *Word normalizer* untuk memperbaiki kata-kata yang terdapat kesalahan ketik, singkatan, dan bahasa yang tidak baku. *Word normalizer* memanfaatkan prosa.ai sebagai sumber NLP API yang dapat mengolah bahasa Indonesia.
4. *Stopword Removal* merupakan proses menghilangkan kata yang umum dan sering muncul dalam teks namun tidak memiliki

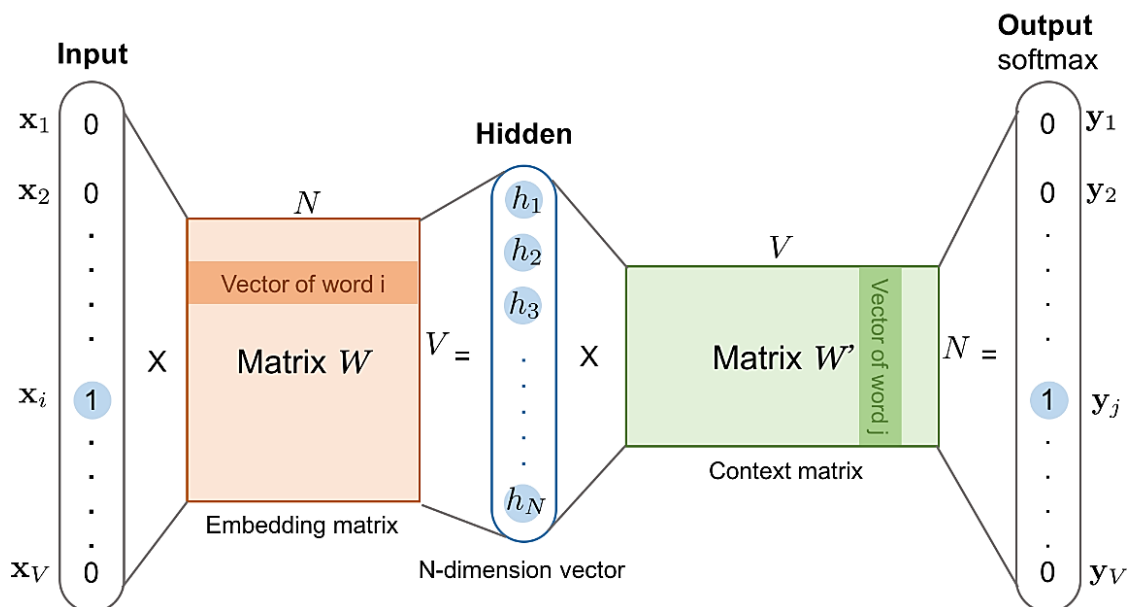
makna seperti kata ganti “saya”, “kamu” atau kata sambung “dan”, “untuk”. Proses ini menggunakan *library* Sastrawi.

5. *Stemming* digunakan untuk mengubah menjadi kata dasar sehingga kata-kata yang memiliki akar yang sama akan dianggap sama. Proses ini menggunakan *library* Sastrawi.

Setelah melalui lima proses fitur NLP diatas, data akan ditokenisasi untuk membagi teks menjadi bagian kecil tiap kata yang akan menjadi inputan pada proses representasi kata dengan word2vec.

2.3. Representasi Kata

Untuk mengidentifikasi representasi vektor analisis sentimen, kami menggunakan model Word2Vec. Word2Vec merupakan model yang efektif dalam mempelajari representasi vektor kata dengan dimensi rendah dari data teks mentah [7]. Variasi Word2Vec yang kami gunakan adalah skip-gram, yaitu model yang mencoba menentukan konteks kata berdasarkan kata-kata di sekitarnya [6]. Kami akan menerapkan hasil dari keempat skenario *preprocessing* data pada model skip-gram Word2Vec dan menguji berbagai nilai dari *size*, *window size*, metode efisiensi komputasi, dan dimensi penyisipan kata untuk mencari kombinasi yang dapat memberikan akurasi terbaik. Arsitektur model skip-gram ditunjukkan pada Gambar 2.

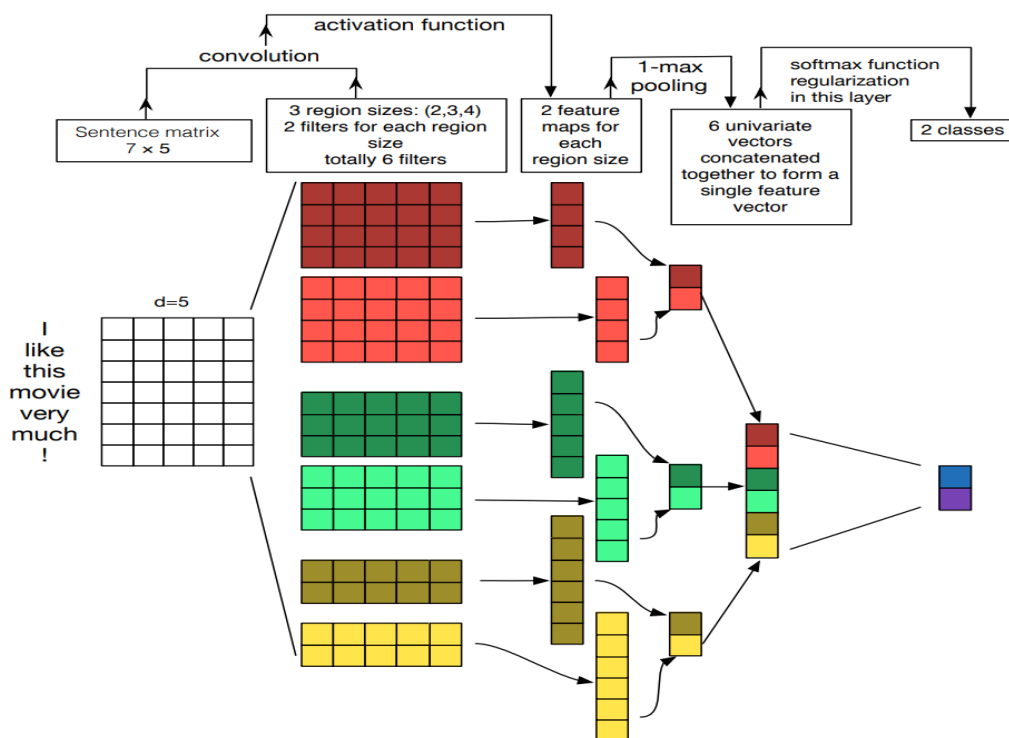


Gambar 2. Arsitektur model Skip-Gram [14]

2.4. Arsitektur Model CNN

Setelah mengubah kata menjadi vektor numerik dengan word2vec, kami akan menggunakan CNN untuk klasifikasi teks dengan memanfaatkan kemampuan jaringan untuk mengekstrak fitur-fitur

penting seperti kata atau frase yang berkaitan dengan topik tertentu [5]. Kami akan mencoba berbagai nilai *kernel size*, *dropout* dan *learning rate* dalam proses pelatihan CNN dan mengevaluasi pengaruhnya terhadap akurasi hasil analisis. Arsitektur Model CNN 1D dapat dilihat pada Gambar 3.



Gambar 3. Arsitektur model CNN 1D [15]

3. HASIL DAN PEMBAHASAN

Setelah melalui tahap pengumpulan data dan *preprocessing*, kami menyusun dataset untuk melakukan analisis sentimen pada ulasan produk. Selanjutnya, kami mencoba berbagai parameter word2vec untuk mengidentifikasi representasi vektor kata yang tepat dan menemukan model yang paling efektif. Kemudian, kami mengaplikasikan hasil dari word2vec pada arsitektur CNN untuk melakukan klasifikasi sentimen pada ulasan produk. Terakhir, kami mengevaluasi berbagai variasi CNN untuk menemukan model yang paling tepat untuk digunakan dengan hasil sebagai berikut.

3.1. Pengumpulan Data

Kami berhasil mengumpulkan 720 produk dengan 20.986 ulasan dari *marketplace* shopee sebagai sumber data utama kami dengan metode scrap. Untuk setiap produk, kami mengambil informasi seperti nama, asal, jumlah terjual, harga, dan url produk. Kemudian untuk setiap ulasan, kami mengambil informasi seperti *username*, isi, tag, tanggal, dan *rating*.

Setelah pengumpulan data, kami melakukan *cleaning* data untuk menghilangkan data yang tidak valid. Kami menghapus data yang memiliki isi kosong atau kesalahan pengambilan isi ulasan, data yang berisi respon dari penjual, data yang tidak memiliki alfabet dalam isi ulasannya, dan data duplikat dari isi ulasan. Setelah proses *cleaning* data, tersisa sebanyak 16.995 ulasan.

Selanjutnya, kami melakukan proses *labeling* data pada sisa data sebanyak 16.995 ulasan untuk mengurangi bias dan subjektivitas dalam pemberian label. Hasil dari proses ini adalah 515 ulasan yang dianggap positif, 490 ulasan yang dianggap negatif, dan 15.990 ulasan yang dianggap tidak relevan. Hanya ulasan yang berlabel positif dan negatif yang akan digunakan dalam proses selanjutnya.

3.2. Preprocessing Data

Dari hasil pengumpulan, pembersihan dan pemberian label pada data, selanjutnya data akan melalui proses pembentukan data yang siap untuk direpresentasikan. Pada tahap *preprocessing* ini kami menggunakan 4 skenario. Fitur *preprocessing* dari keempat skenario dapat dilihat pada tabel 1.

Tabel 1. Skenario fitur *preprocessing*

Skenario	Fitur NLP
1	Case folding, punctuation removal, word normalizer, stopwords removal, dan stemming
2	Case folding, punctuation removal, stopwords removal, dan stemming
3	Case folding, punctuation removal, word normalizer, dan stemming
4	Case folding, punctuation removal, word normalizer, dan stopwords removal

Hasil dari *preprocessing* empat skenario yang dilakukan ditunjukkan pada tabel 2. Contoh ulasan yang diambil adalah salah satu dari dataset yang memiliki ulasan “tipis banget bajunyanembus keliatan. yauda la harganya jg ga worth it”. Tabel 2 juga menampilkan hasil akurasi yang didapatkan dari setiap skenario.

Tabel 2. Hasil *preprocessing* berdasarkan skenario

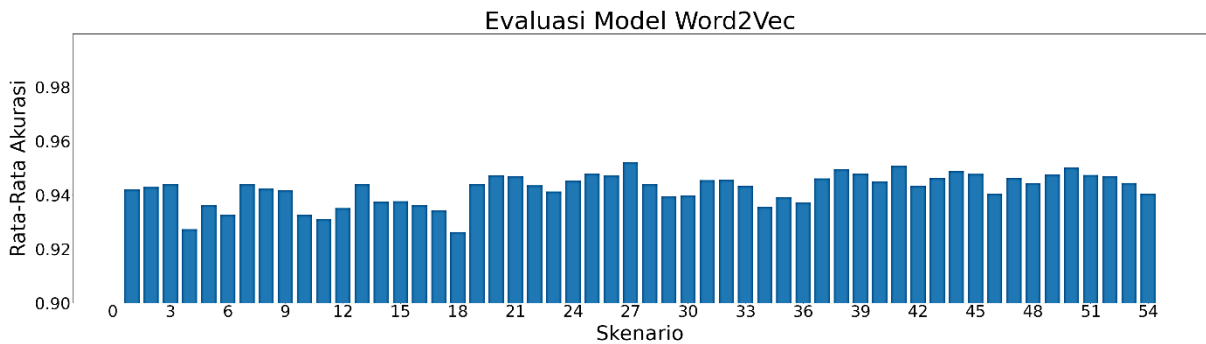
Skenario	Hasil <i>Preprocessing</i>	Akurasi
1	tipis banget baju tembus harga timpal	98.01%
2	tipis banget bajunyanembus liat yauda la harga jg ga worth it	97.01%
3	tipis banget baju tembus lihat ya sudah harga juga enggak timpal	99.00%
4	tipis banget bajunya menembus harganya setimpal	98.51%

Hasil dari masing-masing skenario terhadap model yang telah dibangun dapat dilihat pada tabel 2 yang menunjukkan bahwa skenario ketiga mendapatkan nilai akurasi tertinggi, yaitu sebesar 99,00% dengan nilai presisi 98,96% dan *recall* 98,96%. Di sisi lain, skenario kedua mendapatkan nilai akurasi terendah, yaitu sebesar 97,01% dengan nilai presisi dan *recall* skenario kedua juga lebih rendah dibandingkan skenario ketiga, yaitu masing-masing 95,00% dan 98,96%.

3.3. Word2Vec

Kami menguji 54 skenario Word2Vec dengan 4 pilihan parameter *size* (50, 100, dan 150), *window size* (2, 3, dan 5), metode *hs* (0 dan 1), dan jumlah *negative sampling* (5, 10, dan 20). Setelah menguji semua skenario, kami mengambil parameter Word2Vec yang sama untuk menghitung rata-rata nilai akurasi dari setiap model. Gambar 4 menunjukkan hasil evaluasi dari seluruh skenario Word2Vec, yang menyajikan rata-rata akurasi dari setiap skenario.

Berdasarkan hasil evaluasi tersebut, kami menemukan bahwa kombinasi parameter untuk model word2vec ke 27 (*size* 100, *window* 3, *hs* 0, *negative* 20) memberikan rata-rata yang tertinggi, dengan akurasi sebesar 95.21%, presisi sebesar 94.91%, dan *recall* sebesar 95.13%. Sedangkan kombinasi parameter untuk model word2vec ke 18 (*size* 50, *window* 5, *hs* 1, *negative* 20) memberikan rata-rata yang terendah, dengan akurasi 92.63%, presisi 91.95%, dan *recall* 92.82%.



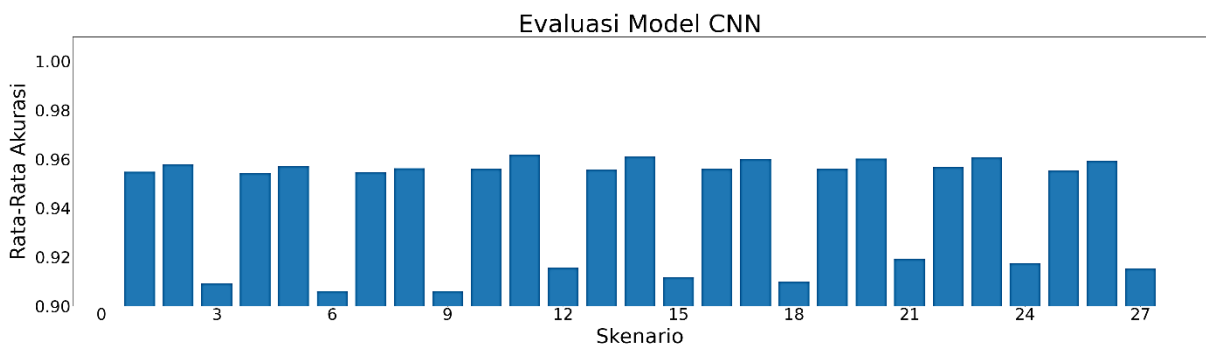
Gambar 4. Rata-rata akurasi model Word2Vec

3.4. CNN

Pada arsitektur CNN kami menguji 3 parameter dengan nilai yang berbeda. 3 pilihan parameter tersebut adalah *kernel size* (2, 3, dan 4), *window size* (0.1, 0.2, dan 0.3), dan *learning rate* (0.01, 0.001, 0.0001) yang menghasilkan 27 model CNN. Kami menguji semua model satu persatu dengan semua hasil *preprocessing* dan semua model representasi kata yang sudah dilakukan serta mengambil model yang sama untuk menghitung nilai rata-rata nilai akurasi dari model CNN. Gambar 5 menunjukkan

hasil evaluasi dari seluruh skenario CNN, yang menyajikan rata-rata akurasi dari setiap skenario.

Pada hasil evaluasi yang ditunjukkan oleh gambar 5, kami menemukan kombinasi parameter ke 11 (*kernel size* 3, *dropout* 0.1, *learning rate* 0.001) memberikan rata-rata yang tertinggi, dengan akurasi sebesar 96.19%, presisi sebesar 95.64%, dan *recall* sebesar 96.44%. Dan model dengan nilai terendah ada pada kombinasi parameter ke 6 (*kernel size* 2, *dropout* 0.2, *learning rate* 0.0001) yang memiliki akurasi sebesar 90.59%, presisi sebesar 89.94%, dan *recall* 90.52%.



Gambar 5. Rata-rata akurasi model CNN

4. DISKUSI

Hasil yang diperoleh dari penelitian ini menunjukkan bahwa kombinasi teknik *preprocessing*, parameter *word2vec*, dan parameter CNN yang kami gunakan mampu menghasilkan model yang sangat akurat dan efektif dalam memprediksi data. Akurasi sebesar 99.00%, presisi 98.96%, dan *recall* 98.96% menunjukkan bahwa model tersebut mampu memprediksi dengan tepat hampir seluruh data yang diberikan.

Kami juga melakukan perbandingan dengan penelitian sebelumnya yang menggunakan metode yang sama, dan hasil yang kami peroleh ternyata lebih baik dibandingkan dengan penelitian tersebut. Penelitian [16] dan [6] masing-masing menganalisis sentimen pada data twitter dan mendapatkan nilai akurasi terbaik berturut-turut sebesar 76,40% dan 88,21%. Dalam penelitian [17] melakukan analisis sentimen terhadap produk pada *marketplace* dalam bahasa china dengan akurasi model CNN yang didapatkan sebesar 90,90%. Penelitian lain [18], model CNN menghasilkan nilai *F1-Score* sangat baik pula sebesar 98,72%.

Pada tabel 2 menunjukkan teknik *word normalizer* sangat membantu dalam meningkatkan akurasi model karena memperbaiki kata yang salah atau ditulis dengan kurang baik oleh pemberi ulasan. Hasil yang menggunakan teknik *preprocessing stopword removal* cenderung memiliki akurasi yang lebih kecil karena menghapus kata yang dapat memberikan makna pada ulasan produk. Berdasarkan hasil evaluasi *word2vec* melalui gambar 4 menunjukkan bahwa kombinasi parameter *word2vec* dengan metode *hs 0* menghasilkan rata-rata akurasi yang lebih tinggi dibandingkan dengan metode *hs 1*. Ini dikarenakan metode *hs 0* menggunakan *negative sampling* yang membantu model memahami konteks kata yang dilatih, sedangkan metode *hs 1* memecah kata ke dalam bagian-bagian kecil sebelum mencari kata terkait yang membutuhkan waktu lebih lama. Hal dapat ini menyebabkan rata-rata akurasi pada metode *hs 1* lebih rendah dibandingkan dengan metode *hs 0*.

Pada gambar 5 memperlihatkan setiap perubahan nilai pada parameter *learning rate* memberi pengaruh signifikan terhadap rata-rata akurasi model. *Learning rate* 0.001 dan 0.01 memiliki pengaruh yang lebih besar dibandingkan dengan *learning rate* 0.0001 yang memperlihatkan penurunan signifikan pada rata-rata akurasi model. *Learning rate* mempengaruhi seberapa cepat model belajar dan menentukan seberapa besar perubahan yang dilakukan terhadap parameter model pada setiap iterasi. Nilai *learning rate* yang lebih tinggi dapat menyebabkan model belajar lebih cepat, namun juga dapat menyebabkan model tidak stabil. Sebaliknya, nilai *learning rate* yang lebih rendah dapat menyebabkan model menjadi lebih stabil, namun juga dapat menyebabkan proses belajar model menjadi lebih lambat.

5. KESIMPULAN

Berdasarkan hasil dari penelitian ini, kombinasi teknik *preprocessing* ketiga (*case folding*, *removal punctuation*, *word normalizer*, dan *stemming*), kombinasi parameter untuk model *word2vec* kedua (*size 50*, *window 2*, *hs 0*, dan *negative 10*), dan kombinasi parameter untuk model CNN keempat (*kernel size 2*, *dropout 0.2*, dan *learning rate 0.01*), terbukti mampu menghasilkan akurasi terbaik dibandingkan dengan kombinasi teknik *preprocessing* dan kombinasi parameter untuk *word2vec* dan CNN lain yang dicoba. Model ini memiliki akurasi 99.00%, presisi 98.96%, dan *recall* 98.96%. Ini menunjukkan kualitas yang sangat baik dan efektif dalam memprediksi data.

Meskipun demikian, ada beberapa kekurangan dari metode yang kami gunakan. Salah satunya adalah waktu yang dibutuhkan dalam melakukan percobaan seluruh skenario dan membangun model membutuhkan waktu yang cukup lama. Namun, kami yakin bahwa penggunaan waktu dapat ditingkatkan. Untuk penelitian selanjutnya, kami akan mencoba mencari teknik dalam mencoba seluruh skenario dan membangun model yang lebih efisien untuk mengurangi waktu yang dibutuhkan.

DAFTAR PUSTAKA

- [1] I. R. Servanda, P. R. K. Sari, dan N. A. Ananda, "Peran Ulasan Produk dan Foto Produk yang Ditampilkan Penjual pada Marketplace Shopee Terhadap Minat Beli Pria dan Wanita," *Jurnal Manajemen dan Bisnis*, vol. 2, hlm. 69–79, 2019.
- [2] M. Nurul, N. Soewarno, dan Isnalita, "Pengaruh Jumlah Pengunjung, Ulasan Produk, Reputasi Toko Dan Status Gold Badge pada Penjualan Dalam Tokopedia," *e-Jurnal Akuntansi*, vol. 28, hlm. 1855–1865, Sep 2019.
- [3] L. S. W. W. Kean, Indriati, dan Marji, "Analisis Sentimen Review Shopee Berbahasa Indonesia Menggunakan Improved K-Nearest Neighbor dan Jaro Winkler Distance," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 3, hlm. 7172–7179, Jul 2019.
- [4] E. H. Muktafin, Kusri, dan E. T. Luthfi, "Analisis Sentimen pada Ulasan Pembelian Produk di Marketplace Shopee Menggunakan Pendekatan Natural Language Processing," *Jurnal Eksplorasi Informatika*, vol. 10, hlm. 32–42, Sep 2020.
- [5] A. H. Ombabi, O. Lazzez, W. Ouarda, dan A. M. Alimi, "Deep Learning Framework based on Word2Vec and CNN for Users Interests Classification," *Sudan Conference on Computer Science and Information Technology (SCCSIT)*, 2017.

- [6] M. R. Aldiansyah dan P. S. Sasongko, "Twitter Sentiment Analysis About Public Opinion on 4G Smartfren Network Services Using Convolutional Neural Network," *International Conference on Informatics and Computational Sciences (ICICoS)*, vol. 3, hlm. 1–6, 2019.
- [7] S. Smetanin dan M. Komarov, "Sentiment Analysis of Product Reviews in Russian using Convolutional Neural Networks," *Conference on Business Informatics (CBI)*, vol. 21, hlm. 482–486, 2019.
- [8] R. S., Arthi.R, S. Murugan, dan Julie, "Topic categorization of Tamil News Articles using PreTrained Word2Vec Embeddings with Convolutional Neural Network," *IEEE International Conference on Computational Intelligence for Smart Power System and Sustainable Energi (CISPSSE)*, hlm. 29–31, Jul 2020.
- [9] A. Nurdin, B. A. S. Aji, A. Bustamin, dan Z. Abidin, "Perbandingan Kinerja Word Embedding Word2vec, Glove, dan Fasttext pada Klasifikasi Teks," *Jurnal TEKNOKOMPAK*, vol. 14, hlm. 74–79, 2020.
- [10] V. Krotov, L. R. Johnson, dan L. Silva, "Tutorial: Legality and Ethics of Web Scraping," *Communications of the Association for Information Systems*, vol. 47, hlm. 555–581, Agu 2020.
- [11] I. R. Hendrawan, E. Utami, dan A. D. Hartanto, "Comparison of Naïve Bayes Algorithm and XGBoost on Local Product Review Text Classification," *Edumatic: Jurnal Pendidikan Informatika*, vol. 6, hlm. 143–149, Jun 2022.
- [12] R. Wagh dan P. Punde, "Survey on Sentiment Analysis using Twitter Dataset," *International conference on Electronics, Communication and Aerospace Technology (ICECA)*, vol. 2, hlm. 208–211, 2018.
- [13] A. N. Rohman, R. L. Musyarofah, E. Utami, dan S. Raharjo, "Natural Language Processing on Marketplace Product Review Sentiment Analysis," *International Conference on Cybernetics and Intelligent System (ICORIS)*, vol. 2, 2020.
- [14] T. Bratanic, "Complete guide to understanding Node2Vec algorithm," *Medium*, 2021. <https://towardsdatascience.com/complete-guide-to-understanding-node2vec-algorithm-4e9a35e5d147> (diakses Des 12, 2022).
- [15] Y. Zhang dan B. C. Wallace, "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification," *The University of Texas*, 2016.
- [16] H. Juwiantho, E. I. Setiawan, J. Santoso, dan M. H. Purnomo, "Sentiment Analysis Twitter Bahasa Indonesia Berbasis Word2vec Menggunakan Deep CNN," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 7, hlm. 181–188, Feb 2020.
- [17] L. Yang, Y. Li, J. Wang, dan R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," *IEEE Access*, vol. 8, hlm. 23522–23530, Jan 2020.
- [18] Y. Luan dan S. Lin, "Research on Text Classification Based on CNN and LSTM," *IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, hlm. 352–355, Mar 2019.