E-ISSN 2621–086X

Indonesian Journal of Applied Statistics

Vol. 5, No. 1, pp. 31-38, May 2022
https://jurnal.uns.ac.id/ijas/article/view/56825
https://doi.org/10.13057/ijas.v5i1.56825

Analisis Sentimen dari Aplikasi Shopee Indonesia Menggunakan Metode Recurrent Neural Network

Herni Utami*
Departemen Matematika, Universitas Gadjah Mada, Yogyakarta, Indonesia

***Corresponding author:** herni_utami@ugm.ac.id

Abstract. *Sentiment analysis on unbalanced data will cause classification errors where the classification results tend to be in the majority class. Therefore, it is necessary to handle unbalanced data. In this study, a combination of synthetic minority oversampling technique (SMOTE) and Tomek link methods will be used to handle unbalanced data. In this study, we use the Recurrent Neural Network (RNN) method to analyze the sentiment of Shopee application users based on review data. Shopee Indonesia application review data shows that around 80% of Shopee application users have positive sentiments and 20% have negative sentiments, which means the data is not balance. In this study, preprocessing process with combination of synthetic minority oversampling technique (SMOTE) and Tomek link method used to handle the condition. The performance of the result is quite good, namely 80% accuracy, 84.1% precision, 92.5% sensitivity, 30% specificity, and 88.1% F1-score. It is better than performance of sentiment analysis that without preprocessing to handle imbalanced data.*

Keywords: *imbalanced data; RNN; sentiment analysis; SMOTE; Tomek link*

1. PENDAHULUAN

Analisis sentimen merupakan salah satu metode untuk mengklasifikasi objek ke dalam dua kategori sentiment positif dan negative. Analisis ini sangat popular dan sering digunakan untuk mengetahui bagaimana respons masyarakat terhadap suatu produk. Dalam analisis sentiment ini, salah satu metode yang dapat digunakan adalah metode *Recurrent Neural Network* (RNN). *Recurrent Neural Network* (RNN) ini tidak membuang informasi begitu saja dari masa lalu, sehingga RNN mampu mengenali pola data dengan baik berdasarkan ingatan sebelumnya. RNN dapat memproses data secara sekuensial. Data sekuensial dapat berupa data teks, data runtun waktu, data suara, dan lain sebagainya. Data yang diperoleh dari media sosial merupakan data teks, yang berupa data sekuensial. Beberapa peneliti telah meneliti metode RNN melakukan analisis sentimen. Diantaranya, Thomas, *et al.* [1] telah mengimplementasikan analisis sentimen untuk data tweet dalam Bahasa Malayalam di India Selatan dengan menggunakan RNN-*Long short term memory* (LSTM) dan teknik *deep learning* untuk memprediksi analisis sentimen. Srividya, *et al.* [2] membandingkan performa klasifikasi sentimen menggunakan LSTM-RNN, Naïve Bayes dan Regresi Logistik, sedangkan Kumalasari, *et al.* [3] meneliti model klasifikasi analisis sentimen menggunakan *deep learning* dan *neural network.*

Data real yang diperoleh dari suatu sumber dapat megandung kelas yang tidak seimbang atau *imbalance*. Data tidak seimbang mengakibatkan kesalahan klasifikasi kelas minoritas karena

Submitted: 29 November 2021; ***Accepted:*** 29 May 2022

data cenderung mendukung kelas mayoritas [4]. Pada kasus riil, terdapat dua kondisi himpunan data dalam klasifikasi, yaitu seimbang dan data tidak seimbang. Kelompok data tidak seimbang merupakan kondisi yang tidak seimbang antara kelas satu dengan kelas lain. Adanya kondisi data tidak seimbang pada analisis klasifikasi dapat memberikan hasil yang tidak optimal.

Untuk penanganan data tidak seimbang perlu adanya *preprocessing*. Teknik *preprocessing* merupakan pendekatan yang mudah dilakukan karena tidak terikat pada metode analis data utama yang digunakan. Teknik ini memodifikasi distribusi data *training* sehingga kedua kelas data (mayoritas dan minoritas) dapat direpresentasikan dengan baik dalam data *training*. Teknik ini dibedakan menjadi dua yaitu metode *oversampling* dan *undersampling*. Metode *oversampling* dilakukan untuk menyeimbangkan jumlah distribusi data dengan cara meningkatkan jumlah data kelas minoritas. Masalah umum yang akan muncul dari metode *oversampling* adalah masalah *overfitting* yang menyebabkan aturan klasifikasi menjadi semakin spesifik meskipun akurasi untuk data *training* semakin membaik. Sedangkan metode *undersampling* dilakukan dengan cara mengurangi banyaknya data pada kelas mayoritas sehingga data menjadi seimbang. Metode ini akan kehilangan informasi dari data yang dihilangkan. Salah satu metode *oversampling* adalah *Synthetic Minority Oversampling Technique* (SMOTE), pertama kali diperkenalkan oleh Chawla, *et al.* [5]. Pendekatan ini bekerja dengan membuat "*synthetic*" data, yaitu data replikasi dari kelas minoritas. Algoritma SMOTE digunakan oleh Chawla pada klasifikasi dengan pohon keputusan. Metode SMOTE merupakan metode yang kuat untuk menangani masalah data tidak seimbang dan telah sukses dalam berbagai macam kasus aplikasi akan tetapi masalah umum yang akan muncul seperti yang disebutkan sebelumnya adalah masalah *overfitting*. Sedang metode *undersampling* yang dapat digunakan salah satunya adalah Tomek Links. Tomek Links diperkenalkan oleh Tomek [6]. Metode ini bekerja dengan menghapus data kelas negatif (mayoritas) yang memiliki kesamaan karakteristik dengan kelas minoritas sehingga data menjadi seimbang. Metode ini akan kehilangan informasi dari data yang dihilangkan. Untuk mengatasi kelemahan metode *oversampling* dan *undersampling*, perlu mengkombinasikan kedua metode ini.

Pada penelitian ini diterapkan metode penanganan data tidak seimbang dengan metode kombinasi SMOTE dan Tomek Links. Sebelumnya banyak penelitian tentang metode *recurrent neural network*, tapi masih jarang untuk data yang tak seimbang dengan penanganan menggunakan metode kombinasi SMOTE dan Tomek Links. Pada penelitian ini mengaplikasikan analisis sentimen pada pengguna aplikasi Shopee Indonesia.

2. METODE KOMBINASI SMOTE DAN TOMEK LINKS

Menurut Vimalraj [7] masalah data tidak seimbang terjadi ketika suatu *database* mempunyai 90% hal berasal dari kelas mayoritas dan sisanya merupakan kelas minoritas. Metode kombinasi SMOTE dan Tomek Links merupakan gabungan dari metode SMOTE dan metode Tomek Links, dengan melakukan secara berurutan dari metode SMOTE yang kemudian dilanjutkan dengan metode Tomek Links sebagai metode pembersihan data.

Algoritma metode kombinasi SMOTE dan Tomek Links sebagai berikut.

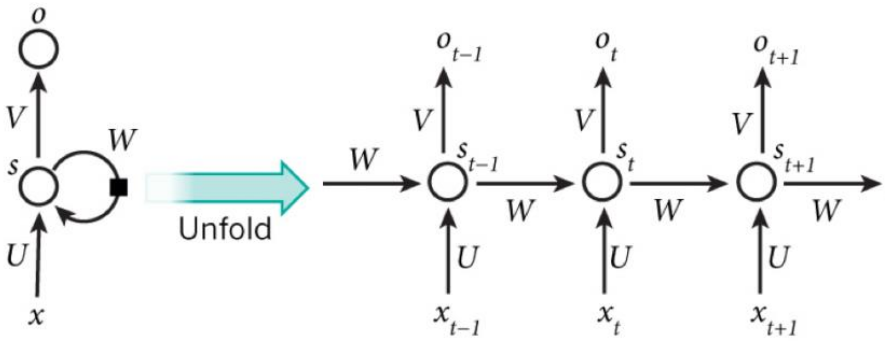
1. Menjalankan algoritma metode SMOTE.
   Langkah ini dimulai dengan menambah jumlah observasi pada kelas minoritas dengan membuat objek atau observasi sintetis, yaitu objek baru yang tidak terdapat dalam dataset namun memiliki kemiripan dengan objek yang terdapat dalam dataset.

Observasi sintetis dibentuk dari dua observasi, dengan observasi pertama dipilih dari data kelas minoritas dan observasi kedua dari data kelas minoritas yang dipilih secara random dengan k-nearest neighbor observasi kelas minoritas yang pertama. Dengan adanya observasi sintetis tersebut maka jumlah observasi pada data kelas minoritas akan bertambah sehingga lebih seimbang dengan data kelas mayoritas.

2. Identifikasi Tomek Links pada data hasil SMOTE.
   Sepasang observasi disebut sebagai Tomek Links apabila kedua observasi tersebut merupakan tetangga terdekat namun memiliki kelas yang berbeda.
3. Pasangan observasi yang teridentifikasi sebagai Tomek Links dihapus dari dataset.
   Melakukan pengulangan identifikasi Tomek Links hingga menghasilkan data yang bersih dari *noise*.

3. *RECURRENT NEURAL NETWORK* (RNN)

Menurut DiPietro dan Hager [8] RNN adalah bagian dari *Neural Network* untuk memproses data sekuensial atau data yang mempunyai urutan. RNN akan menyimpan informasi data masa lalu untuk mengetahui pola data. Proses kerja RNN terlihat pada Gambar 1.



Gambar 1. Arsitektur *Recurrent Neural Network*
(Sumber: machinelearning.mipa.ugm.ac.id)

Hidden state (s_t) dan output (o_t) pada RNN untuk langkah ke *t* diformulasikan dengan

$$s_t = f(Ux_t + Ws_{t-1}) \quad o_t = softmax(Vs_t)$$

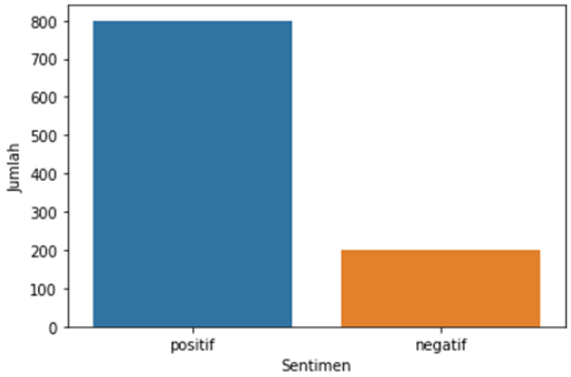
dengan *U, W, V* merupakan parameter bobot dalam proses RNN secara berurut, yaitu parameter antara input dan *hidden state*, parameter antar *hidden state*, dan parameter antara *hidden state* dan *output*.

4. HASIL DAN PEMBAHASAN

Pengambilan data dilakukan menggunakan bahasa pemrograman Python dengan *library* google-play-scraper dimana *library* tersebut dapat memberikan akses ke API twitter, sehingga dapat dilakukan pengambilan data dengan mudah di Google Play Store. Data yang digunakan untuk analisis sentimen merupakan *review* dari Aplikasi Shopee Indonesia. Shopee merupakan aplikasi yang digunakan untuk jual beli secara *online*. *Review* yang diperoleh berupa ulasan dalam teks berbahasa Indonesia dan rating berupa bintang 1 sampai 5. Data yang diperoleh dari Google Play Store tidak memiliki label kelas sentimen. Menurut Nguyen [9,10], rating dapat digunakan untuk memberikan label. Rating dengan bintang 1, 2, dan 3 dapat diberi label ‘negatif’, sedangkan rating dengan bintang 4 dan 5 dapat diberi label positif.

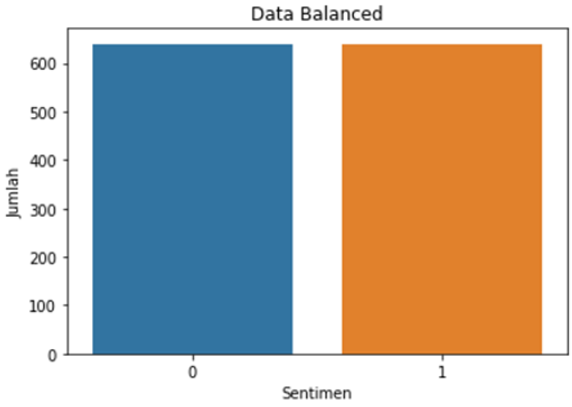
Berdasarkan 1.000 *review* pengguna Aplikasi Shopee yang disajikan pada Gambar 2, terlihat 800 *review* mempunyai sentimen positif dan 200 *review* mempunyai sentimen negatif. Selanjutnya dilakukan *preprocessing* data sebagai berikut.

1. Menghapus duplikasi data
   Duplikasi data merupakan data yang dituliskan dengan sama persis lebih dari satu kali.
2. *Case folding*
   Pada tahap *case folding* dilakukan pengubahan semua huruf kapital (*uppercase*) menjadi huruf kecil (*lowercase*). Perbedaan penggunaan huruf kapital dapat mempengaruhi dalam proses analisis karena dapat dianggap sebagai kata yang berbeda, sehingga seluruh kata perlu diubah menjadi *lowercase*.
3. *Cleansing*
   Cleansing merupakan penghapusan karakter selain huruf yang dianggap sebagai *delimiter* (*separator*) dan karakter spesial. Pada tahap ini dilakukan *remove punctuation*, yaitu akan dilakukan proses menghilangkan simbol atau tanda baca yang dapat mengganggu proses analisis. Pada tahap ini, tanda baca yang dihapus seperti ~!@#$%^&*()-=[]<>. Selain itu, juga dilakukan penghapusan nomor, *whitespace* dan URL.
4. *Stop word removal*
   Stop word removal merupakan tahap untuk menghapus kata-kata yang tidak memiliki makna terhadap teks. Seperti kata penghubung, diantaranya “dari”, “ke”, dan sebagainya. Penghapusan *stop word* menyisakan kata-kata penting yang akan diproses ke tahap selanjutnya.



Gambar 2. *Rating review* pengguna Aplikasi Shopee

Pada tahap analisis data digunakan data *training* dan *testing* dengan perbandingan 80:20. Sehingga, digunakan 800 *review* pada data *training* dengan 640 untuk sentimen positif dan 160 untuk sentimen negatif. Untuk melakukan penanganan data tidak seimbang dengan metode kombinasi SMOTE dan Tomek Links maka dalam kasus ini, data sentimen yang berupa teks, terlebih dahulu diubah menjadi numerik, yaitu dengan melakukan pembobotan. *Term Frequency-Inverse Document Frequency* atau TF-IDF merupakan salah satu metode algoritma yang berguna untuk menghitung bobot setiap kata. Setelah dilakukan penanganan data tidak seimbang diperoleh hasil 640 data untuk sentimen positif dan 640 data untuk sentimen negatif seperti tampak pada Gambar 3.



Gambar 3. Data *review* setelah ditangani dengan kombinasi SMOTE dan Tomek Links

Pada penelitian ini, analisis sentimen dengan RNN dilakukan untuk data asli (tak seimbang) dan juga untuk data setelah ditangani ketidakseimbangannya dengan metode kombinasi SMOTE dan Tomek Links. Hasil analisis akan dibandingkan. Pertama dilakukan analisis sentimen dengan RNN untuk data asli.

Tabel 1. Ringkasan model RNN berdasarkan data asli

Layer (Type)	*Output Shape*	# Parameter
Embedding_7 (Embedding)	(None, 83, 64)	184.576
Simple_rnn_7 (SimpleRNN)	(None, 64)	8.256
Dense_7 (Dense)	(None, 2)	130

Total parameter: 192.962
Trainable parameter: 192.962
Non-trainable parameter:0

Tabel 1 menunjukkan informasi tipe layer, *output shape*, dan banyaknya parameter yang digunakan. Analisis sentimen dilakukan dengan metode RNN dengan dimensi *word embedding* 64 dan panjang ukuran input sebesar 83, yaitu jumlah kata terbanyak dari *review* seluruh data *training*. Jumlah *neuron* RNN 64, dan *output* diklasifikasikan ke dalam 2 kelas. Terdapat 192.962 parameter yang dilatih dan tidak ada parameter yang tidak dilatih. Fungsi aktivasi tanh yang digunakan pada *hidden state* adalah fungsi tanh dan fungsi aktivasi pada *output layer* adalah *softmax*. Serta digunakan *Binary Cross Entropy Loss Function*, optimasi Adam, dan *epoch* sebesar 10. Berdasarkan hasil analisis, diperoleh hasil seperti Gambar 4.

Dengan menggunakan Model *Checkpoint* pada *Keras Library* untuk menyimpan model atau bobot model yang terbaik akan menghasilkan nilai *test loss* yang terendah. Terlihat bahwa nilai *loss* terendah dihasilkan pada *epoch* kesembilan sehingga model beserta bobot pada *epoch* kesembilan akan disimpan dan dapat digunakan untuk prediksi. Hasil prediksi ditunjukkan dengan *Confussion Matrix* yang disajikan pada Tabel 2.

Gambar 4. Grafik akurasi dan *loss model* untuk data *rating review* Shopee

Tabel 2. *Confussion matrix*

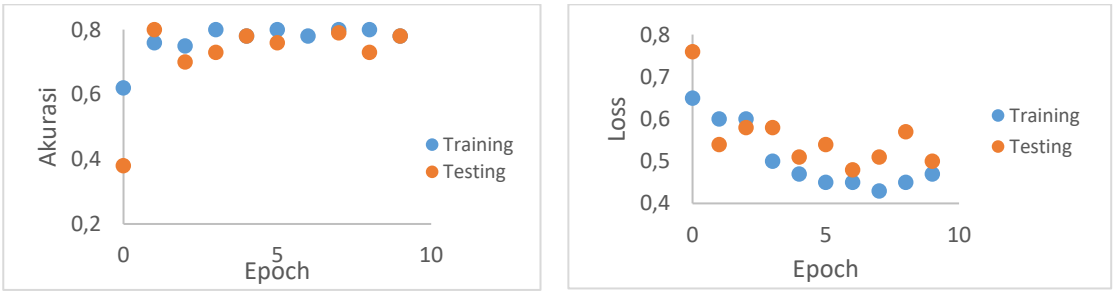
Aktual dan prediksi		Prediksi	
		Positif	Negatif
Aktual	Positif	148	12
	Negatif	39	1

Analisis sentimen kedua dilakukan berdasarkan data *rating review* pengguna Shopee yang sudah seimbang setelah ditangani menggunakan metode kombinasi SMOTE dan Tomek links. Ringkasan model RNN hasil analisis data *rating review* disajikan pada Tabel 3.

Tabel 3. Ringkasan model RNN berdasarkan data yang sudah seimbang

Layer (Type)	*Output Shape*	# Parameter
Embedding_9 (Embedding)	(None, 83, 64)	184.576
Simple_rnn_9 (SimpleRNN)	(None, 64)	8.256
Dense_9 (Dense)	(None, 2)	130

Total parameter: 192.962
Trainable parameter: 192.962
Non-trainable parameter:0



Gambar 5. Grafik akurasi dan *loss model* untuk data *rating review* Shopee yang sudah seimbang

Tabel 3 menunjukkan model *summary* dari metode RNN dengan dimensi *word embedding* 64 dan panjang ukuran input sebesar 83, yaitu jumlah kata terbanyak dari *review* seluruh data *training*. Jumlah neuron RNN 64, dan *output* diklasifikasikan ke dalam 2 kelas. Terdapat 192.962 parameter yang dilatih dan tidak ada parameter yang tidak dilatih. Pada analisis ini, digunakan

fungsi aktivasi tanh pada *hidden state* dan fungsi aktivasi *softmax* pada *output layer*. Serta digunakan *Binary Cross Entropy Loss Function*, optimasi Adam, dan *epoch* sebesar 10. Hasil analisis ditunjukkan pada Gambar 5.

Berdasarkan model terbaik yang diperoleh yaitu model dengan bobot terbaik pada *epoch* kesembilan disimpan dan dapat digunakan untuk prediksi. Hasil prediksi dapat dilihat pada *confussion matrix*. Tabel 4 merupakan perbandingan rata-rata performa data uji yang dilatih dengan RNN untuk data *review* original (tak seimbang) dan data *review* yang ditangani dengan teknik SMOTE dan Tomek Links (data seimbang).

Tabel 4. *Confussion matrix*

Aktual dan prediksi		Prediksi	
		Positif	Negatif
Aktual	Positif	148	12
	Negatif	28	12

Tabel 5. Perbandingan performa model

Ukuran Performa	Tak Seimbang	Seimbang
Akurasi	0,745	0,800
Presisi	0,791	0,841
Recall	0,925	0,925
Spesifisitas	0,025	0,300
F-1 score	0,853	0,881

Berdasarkan hasil pada Tabel 5, teknik SMOTE dan Tomek Links secara umum mampu meningkatkan performa model untuk mengklasifikasikan pengguna Aplikasi Shopee. Hal ini dapat terlihat adanya peningkatan akurasi, presisi, spesifitas dan *F1-score*. Data tak seimbang dapat menyebabkan *accuracy paradox* dimana hasil klasifikasi diperoleh akurasi yang tinggi tetapi hasil tersebut bias.

5. KESIMPULAN

Data *review* pengguna Aplikasi Shopee Indonesia menunjukkan data yang tidak seimbang dikarenakan sekitar 80% pengguna Aplikasi Shopee di Indonesia mempunyai sentimen positif sedang 20% mempunyai sentimen negative. Dengan menggunakan metode kombinasi *synthetic minority oversampling technique* (SMOTE) dan Tomek Link untuk menangani data yang tidak seimbang ini, maka hasil analisis sentimen menjadi lebih baik dibanding dengan yang tanpa penanganan data tak seimbang. Performa hasil analisis sentimen untuk data pengguna Shopee Indonesia cukup baik yaitu: tingkat akurasi prediksi klasifikasi 80%, presisi 84,4%, sensitivitas 30% specificity, dan F1-Score 88,1%.

DAFTAR PUSTAKA

[1] M. Thomas, and C. A Latha, “Sentimental Analysis Using Recurrent Neural Network”, Int. J. of Eng. and Tech., vol. 7 no. 2.27, pp. 88-92, 2018.

[2] K. Srividya and A. M. Sowjanya, “Aspect Based Sentiment Analysis Using RNN-LSTM”, Int. J. of Adv. Sci. and Tech., vol. 29 no. 04, pp. 5875-5880, 2020.

[3] L. Kumalasari, L. and A. Setyanto, "Sentiment Analysis Using Recurrent Neural Network", J. of Phys.: Conf. Ser., vol. 1471, 012018, pp. 1-6, 2020, doi:10.1088/1742-6596/1471/1/012018

[4] F. Provost, "Machine Learning from Imbalanced data Sets", IAAA Technical Report, 2000.

[5] N.V. Chawla, "SMOTE: Synthetic Minority Oversampling Technique", J. of Art. Intelligence Research, vol. 16, pp. 321-357, 2002.

[6] I. Tomek, "Two modifications of CNN", IEEE Trans. Systems, Man and Cybernetics, vol. 6, pp. 769–772, 1976.

[7] S. Vimalraj and R. Parkodi, "A Review on Handling Imbalanced Data", Proceeding of 2018 IEEE International Conference on Current Trend Toward Converging Technologies, Coimbatore, India, 2018.

[8] R. DiPietro and G. D. Hager, Handbook of Medical Image Computing and Computer Assisted Internation, 503-5014, Academic Press, 2014.

[9] T. L Nguyen, "A Fuzzy Convolutional Neural network for Text Sentiment Analysis", J. of Intelligent and Fuzzy System, vol. 35 no. 6, pp. 6025-6034, 2018.

[10] Y. Mejova, Sentiment Analysis: An Overview Comprehensive Exam Paper. Computer Science Department, University of Iowa, pp. 1-34, 2009.