# Analisis Sentimen dari Aplikasi Shopee Indonesia Menggunakan Metode Recurrent Neural Network

*by* Herni Utami

**Submission date:** 27-May-2022 02:18PM (UTC+0700)
**Submission ID:** 1845185266
**File name:** Vol5_No1_Utami_5_-.pdf (250.57K)
**Word count:** 2289
**Character count:** 14148

# Analisis Sentimen dari Aplikasi Shopee Indonesia Menggunakan Metode Recurrent Neural Network

**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 resutl is quite good, namely 80% accuracy, 84.1% precision, 92.5% sensitifity, 30% specificity, and F1-score and 88.1% F1-score.

**Keywords :** sentiment analysis, imbalanced data, tomek link, SMOTE, RNN.

## 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 bisa 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 sentiment. Diantaranya, Thomas and Latha (2018) 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 sentiment. Srividya and Sowjanya (2020) membandingkan performa klasifikasi sentimen menggunakan LSTM-RNN, naïve Bayes dan regresi logistic. Sedangkan Kumalasari dan Setyanto (2020) 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 data cenderung mendukung kelas mayoritas (Provost, 2000). 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 *prepropcessing*. 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 jumlah data 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 dkk (2002). 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 (1976). 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 akan di terapkan 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 sentiment pada pengguna aplikasi Shopee Indonesia.

**2. Metode Kombinasi SMOTE dan Tomek Links**

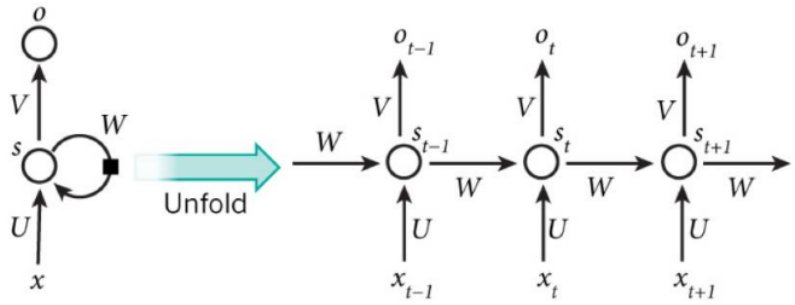
Menurut Vimalraj (2018) masalah data tidak seimbang terjadi ketika suatu database mempunyai 90% hal berasal dari kelas mayoritas dan sisanya merupakan kelas minoritas. Salah satu metode yang dapat digunakan untuk menangani data tak seimbang adalah metode kombinasi SMOTE dan Tomek Links. Metode kombinasi 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 adalah 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*.
4. Melakukan pengulangan identifikasi Tomek Links hingga menghasilkan data yang bersih dari *noise*.

**3. Recurrent Neural Network (RNN)**

Menurut DiPietro dan Hager (2020) 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 berikut.




Gambar 1. Arsitektur Recurrent Neural Network

(machinelearning.mipa.ugm.ac.id)

Hidden state ($s_t$) dan output $(o_t)$ pada RNN untuk langkah ke t diformulasikan sebagai berikut:

$$s_t = f(Ux_t + Ws_{t-1})$$

$$o_t = softmax(Vs_t)$$

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

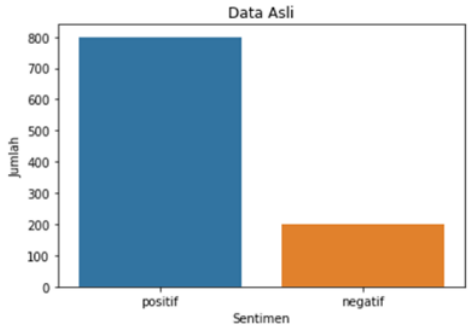
## 4. Hasil dan Diskusi

### 4.1 Pengumpulan Data

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 dkk, 2018), 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.




Gambar 2. *Rating review*

Berdasarkan 1.000 review pengguna aplikasi shopee yang di sajikan 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 penggunakaan huruf kapital dapat mempengaruhi dalam proses analisis karena dapat dianggap sebagai kata yang berbeda. Sehingga, seluruh kata perlu diubah menjadi *lowercase*.

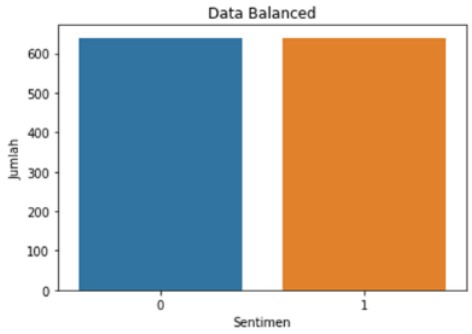
3. *Cleansing*

   *Cleansing* merupakan penghapusan karakter selain huruf yang dianggap sebagai delimiter (separator) dan karakter spesial. Pada tahap ini akan 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 menghilangan kata-kata yang tidak memiliki makna terhadap teks. Seperti kata pengubung “dari”, “ke”, dan sebagainya. Penghapusan stop word menyisakan kata-kata penting yang akan diproses ke tahap selanjutnya.

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 negative seperti tampak pada Gambar 3.




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

Pada penelitian kali ini, analisis sentiment dengan RNN akan di lakukan untuk data asli (tak seimbang) dan juga untuk data setelah ditangani ketidakseimbangannya dengan metode kombinasi SMOTE dan tomek links. Hasil analisis akan dibandingkan. Pertama akan dilakukan analisis sentiment dengan RNN untuk data asli.

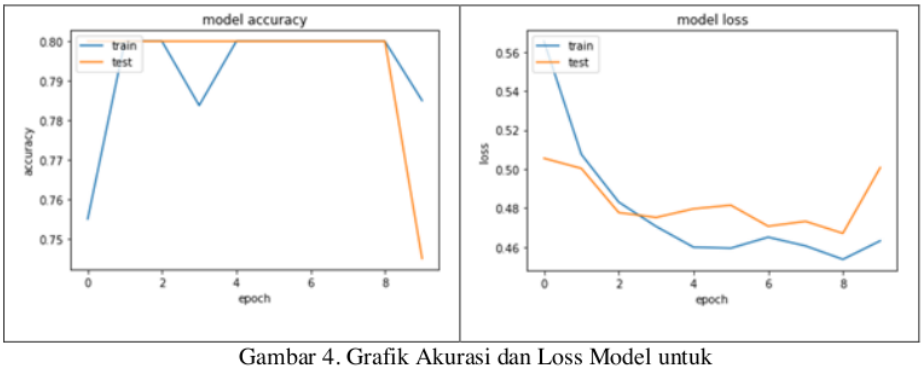
Tabel 1. Ringkasan model RNN berdasarkan data asli

| Layer (type) | Output Shape | Param # |
|---|---|---|
| embedding_7 (Embedding) | (None, 83, 64) | 184576 |
| simple_rnn_7 (SimpleRNN) | (None, 64) | 8256 |
| dense_7 (Dense) | (None, 2) | 130 |

Total params: 192,962
Trainable params: 192,962
Non-trainable params: 0

None

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 sebagai berikut:




Gambar 4. Grafik Akurasi dan Loss Model untuk data rating review Shopee

Dengan menggunakan *ModelCheckpoint* pada *keras* 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 dapat dilihat pada confussion matrix berikut:

Tabel 2. Matriks Confussion

| | | Prediksi | |
|---|---|---|---|
| | | Positif | Negatif |
| Aktual | Positif | 148 | 12 |
| | Negatif | 39 | 1 |

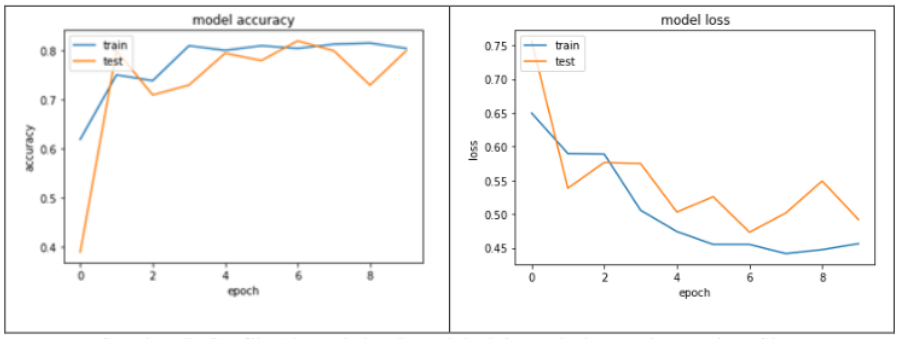
Analisis sentiment kedua akan dilakukan berdasarkan data *rating review* pengguna Shopee yang sudah seimbang setelah di tangani menggunakan metode kombinasi SMOTE dan Tomek links. Ringkasan model RNN hasil analisis data *rating review* disajikan pada Tabel 3 di bawah ini.

Tabel 3. Ringkasan model RNN berdasarkan data yang sudah seimbang

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_9 (Embedding)      (None, 83, 64)            184576
_________________________________________________________________
simple_rnn_9 (SimpleRNN)     (None, 64)                8256
_________________________________________________________________
dense_9 (Dense)              (None, 2)                 130
=================================================================
Total params: 192,962
Trainable params: 192,962
Non-trainable params: 0
_________________________________________________________________
None
```

Dari Tabel 3 di atas 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. Berdasarkan hasil analisis, diperoleh hasil sebagai berikut:




Gambar 5. Grafik Akurasi dan Loss Model untuk data rating review Shopee yang sudah seimbang

Berdasarkan model terbaik yang diperoleh yaitu model dengan bobot terbaik pada *epoch* kesembilan akan disimpan dan dapat digunakan untuk prediksi. Hasil prediksi dapat dilihat pada confussion matrix berikut:

Tabel 4. Confussion Matrix

| | | Prediksi | |
|---|---|---|---|
| | | Positif | Negatif |
| Aktual | Positif | 148 | 12 |
| | Negatif | 28 | 12 |

Tabel 5 di atas 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 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 di atas, teknik SMOTE dan Tomek Links secara umum mampu meningkatkan performa model untuk mengklasifikasikan pengguna aplikasi Shopee. Hal ini bisa terlihat adanya peningkatan akurasi, presisi, spesifikasi dan F1-score. Data tak seimbang dapat menyebabkan *accuracy paradox* (Zhou, 2007) dimana hasil klasifikasi diperoleh akurasi yang tinggi tetapi hasil tersebut bias.

## 5. Kesimpulan

Data tak seimbang dapat menyebabkan *accuracy paradox* dimana hasil klasifikasi diperoleh akurasi yang tinggi tetapi hasil tersebut bias. Metode kombinasi SMOTE dan Tomek links dapat menangani dengan baik untuk data tak seimbang tentang review pengguna aplikasi shopee. Dengan melakukan *preprocessing* data sebelum dilakukan analisis klasifikasi dapat meningkakan performa model klasifikasi.

## Daftar Pustaka


[1] Chawla, N.V., dkk, SMOTE: Synthetic Minority Oversampling Technique, *Journal of Artificial Intelligence Research*, 16, 321-357, 2002.

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

[3] Kumalasari, L., dan Setyanto, A.,Sentoment Analysis Using Recurrent Neural Network, Journal of Physics Conference Series, 1471(1), 2020.

[4] Nguyen, T.L., dkk, A Fuzzy Convolutional Neural network for Text Sentiment Analysis, Journal of Intelligent and Fuzzy System, 35(6), 6025-6034, 2018.

[5] Provost, F., Machine Learning From Imbalanced data Sets, IAAA technical Report, 2000.

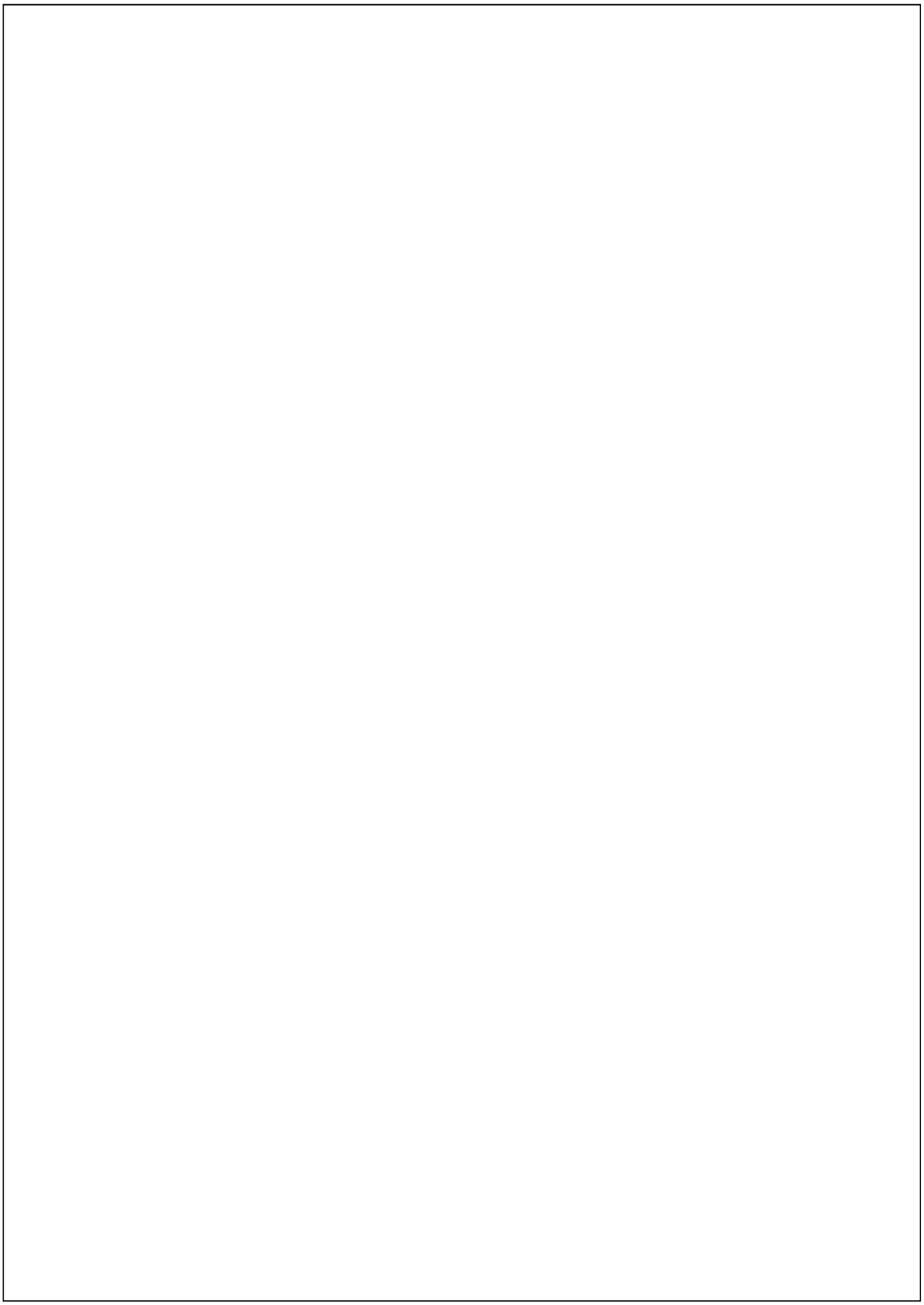
[6] Srividya, K., dan Sowjanya, A.M., Aspect Based Sentiment Analysis Using RNN-LSTM, International Journal of Advanced Science and Technology, 29(04), 2020

[7] Thomas, M., and Latha, C.A., Sentimental Analysis Using Recurrent Neural Network, International Journal of Engineering and Technology, 7(2.27), 88-92, 2018.

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

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

[10] Zhou, dkk, Solving The Apparent Diversity-Accuracy Dilemma of Recommender System, Proceeding of The National Academy of Science of The United States of America, 107(100), 4511-4515, 2010.



# Analisis Sentimen dari Aplikasi Shopee Indonesia Menggunakan Metode Recurrent Neural Network

ORIGINALITY REPORT

| 17% | 15% | 2% | 4% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | repository.its.ac.id<br>Internet Source | 10% |
| 2 | "Data Science", Springer Science and Business Media LLC, 2018<br>Publication | 1% |
| 3 | medium.com<br>Internet Source | 1% |
| 4 | journal.untar.ac.id<br>Internet Source | 1% |
| 5 | Submitted to UIN Sultan Syarif Kasim Riau<br>Student Paper | 1% |
| 6 | Submitted to Universitas Islam Bandung<br>Student Paper | 1% |
| 7 | e-journal.uajy.ac.id<br>Internet Source | 1% |
| 8 | jurnal.unej.ac.id<br>Internet Source | <1% |
| | ojs.uajy.ac.id | |

| | | |
|---|---|---|
| 9 | Internet Source | <1% |
| 10 | tel.archives-ouvertes.fr<br>Internet Source | <1% |
| 11 | www.coursehero.com<br>Internet Source | <1% |
| 12 | portal-uang.com<br>Internet Source | <1% |
| 13 | pt.scribd.com<br>Internet Source | <1% |

| | | | |
|---|---|---|---|
| Exclude quotes | Off | Exclude matches | Off |
| Exclude bibliography | On | | |