Jurnal Khazanah Volume 15 No. 2, November 2023
E-ISSN 2745-8733
P-ISSN1979-519X

AUTOMATIC QURANIC ARABIC COMPLETION FOR QURAN LANGUAGE BASED ON RECURRENT NEURAL NETWORK

Risa Riski Amalia[1], Asti Dwi Sripamuji[2,] Agi Prasetiadi[3*)]
[1,2,3] Institut Teknologi Telkom Purwokerto
Jl. D.I Panjaitan No. 128 Purwokerto 53147, Jawa Tengah - Indonesia

ABSTRAK

Dalam aksara arab semua kata eksplisit bisa dibaca, sedangkan dalam dialek arab bukan tata bahasa yang mudah untuk dimengerti. Sekalipun seseorang dapat menghafal banyak kosakata bahasa Arab (*mufradat*/*vocabulary*) atau menguasai ilmu Tasrif (bukan hanya teori), tidak ada jaminan seseorang tersebut akan bisa membaca dengan tepat teks gundul. *Automatic Quranic Arabic* ini merupakan suatu inovasi dalam bidang Deep Learning yang memberikan *output* berupa harakat secara otomatis untuk teks bahasa Arab gundul. Penelitian sebelumnya menggunakan Long Short-Term Memory (LSTM) untuk memperbaiki kesalahan pada tingkat karakter mendapatkan akurasi 83,76%. Sedangkan pada penelitian ini, kami menawarkan suatu inovasi berupa pengimplementasian algoritma Recurrent Neural Network (RNN) untuk menormalisasikan kembali aksara arab tanpa harakat menjadi tulisan arab yang lengkap dengan harakat. Model kami adalah model pertama di dunia yang mencoba memulihkan kata-kata Bahasa Arab yang sepenuhnya tanpa harakat dengan kalimat yang sepenuhnya berharokat menggunakan dataset training berbasis al-Qur'an dengan pencapaian akurasi 98% komponen arsitektur model LSTM yaitu 3 layer (64, 128, 64), optimizer “adam”, dan activation “relu”.

Kata kunci: Al-Qur’an, Pelengkap Harakat Arab, Long Short-Term Memory (LSTM)

ABSTRACT

In the Arabic script, all explicit words can be read, while in Arabic dialects it is not easy grammar to understand. Even if one can memorize a lot of Arabic vocabulary (mufradat/vocabulary) or master Tasrif science (not just theory), there is no guarantee that a person will be able to read exactly the bare text. Automatic Quranic Arabic is an innovation in the field of Deep Learning that provides output in the form of harakat automatically for plain Arabic texts. Previous research using Long Short-Term Memory (LSTM) to correct errors at the character level gained an accuracy of 83.76%. Meanwhile, in this study, we offer innovation in the form of implementing the Recurrent Neural Network (RNN) algorithm to normalize the Arabic script without harakat into Arabic writing complete with harakat. Our model is the first model in the world to attempt to recover completely devoid Arabic words with fully memorized sentences using a Qur'an-based training dataset with 98% accuracy and the LSTM model architecture components of 3 layers (64, 128, 64), "adam", and activation "relu".

Keywords: *Al-Qur’an, Complementary Arab Harakat, Long Short-Term Memory (LSTM)*

1. PENDAHULUAN

Bahasa Arab berperan penting dalam kehidupan seorang muslim (1). Penggunaan bahasa arab dalam sebuah kitab kuning sering menggunakan tulisan tanpa harakat. Pembuatan model pelengkap harakat ini terinspirasi dari kesusahan pembaca pemula teks bahasa Arab dalam sebuah kitab kuning yang sering menggunakan tulisan tanpa harakat (2-5). Aksara Arab memiliki banyak *diakritik*, termasuk (*i'jam*,إِعْجَام),

konsonan menunjuk, dan *tashkil* (تَشْكِيل,tashkīl), *diakritik* tambahan yang terakhir termasuk tanda vokal ḥarakāt (حَرَكَات) tanda vocal - tunggal: ḥarakah(حَرَكَة) (6-8).

Dalam aksara arab semua kata eksplisit bisa dibaca, sedangkan dalam dialek arab bukan tata yang mudah untuk dimengerti. Padahal untuk membaca tulisan Arab gundul, dibutuhkan penguasaan pemahaman tersendiri tentang ilmu shorof (9). Tidak ada jaminan seseorang dapat membaca tulisan arab gundul walau di sudah menguasai ilmu tasrif dan sudah menghafal banyak kosa kata bahasa arab (10). Maka dari itu, kami membuat model *Automatic Quranic Arabic* untuk memberikan *output* cara baca tulisan Arab gundul tadi dengan menambahkan harakat secara otomatis.

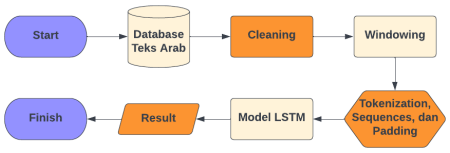
Penelitian ini memberikan suatu inovasi baru berupa pengimplementasian algoritma Recurrent Neural Network (RNN) untuk menormalisasikan kembali aksara arab tanpa harakat menjadi tulisan arab yang lengkap dengan harakat secara otomatis. Hal ini bertujuan untuk membantu pembaca pemula teks bahasa Arab dalam melafalkan tulisannya. Jantung utama model ini adalah sebuah model Deep Learning yang dibangun di atas library Tensorflow, dengan arsitektur otak pilihan berbasis Recurrent Neural Network (RNN), yang sudah kami bangun dan latih dari dataset Al-Qur'an.

RNN adalah sebuah jaringan syaraf tiruan khusus yang dibangun untuk melakukan prediksi dari inputan yang bertipe sekuensial. RNN baik digunakan pada penyelesaian masalah yang menggunakan dataset berupa data terurut. Dalam bahasa sederhananya, model otak yang kami buat, melakukan penerjemahan kalimat dari satu jenis bahasa atau tulisan, ke bahasa atau tulisan yang lain. Bahasa inputnya adalah teks Arab gundul, di mana teks-teks berbahasa Arab tanpa harakat kami jadikan sebagai bahasa asal. Adapun bahasa outputnya adalah teks Arab yang sudah berharakat. Agar otak yang dibuat dapat melakukan penerjemahan antar jenis teks, maka perlu dilakukan fase training di mana otak RNN kami akan mencoba mempelajari hubungan dan pattern antar dua bahasa. Setelah fase training selesai, maka otak siap digunakan dan siap ditugasi untuk menerjemahkan teks input sesuka hati pengguna. Sementara itu, arsitektur yang digunakan yaitu Long Short Term Memory (LSTM) dimana arsitektur ini merupakan modifikasi RNN yang populer dan banyak digunakan dalam program penerjemah bahasa.

Model kami adalah model pertama di dunia yang mencoba memulihkan kata-kata Bahasa Arab yang sepenuhnya tanpa harakat dengan kalimat yang sepenuhnya berharakat dengan kombinasi *epoch*, *step per epoch* , dan menggunakan arsitektur LSTM model seperti berikut 3 layer (64, 128, 64), *optimizer "adam", dan activation "relu",* sehingga mendapatkan hasil akurasi mencapai 98%.

2. METODE

Metode untuk Automatic Quranic Arabic for Qur'an Language Based on Recurrent Neural Network terlihat pada Gambar 2.1 di bawah ini.



Gambar 2.1 *Flowcart Research*

2.1. Dataset ASCII Arabic

Dataset penelitian ini didapatkan dari *Quranic Arabic Corpus (morphology, version 0.4)* ditandai dengan data morfologis yang terkandung dalam *file* teks yang diunduh dari *License* GNU *General Public License*. Mereka menyediakan skrip ASCII *Quranic Arabic* yang mencakup aspek sintaks dan morfologis *Annotation of the Qur'an* dengan teks Arab klasik buatan *Tanziil Project* (11).

2.2. Data *Cleaning*

Dataset yang terakumulasi selama beberapa tahun cenderung tidak lengkap, tidak konsisten dan mengandung data yang bising, data

tersebut akan menyebabkan ketidakkonsistenan di dalam sebuah dataset. Proses cleaning dilakukan untuk mendapatkan data corpus tanpa vocal menggunakan *Regular Expression.* Data hasil *cleaning* ini digunakan sebagai data tanpa kharakat.

2.3. *Windowing* Dataset

Pada proses *windowing* dilakukan pengelompokkan data menjadi tujuh kata di setiap baris, sehingga didapatkan hasil seperti pada gambar 2.4

0	قوله (العلم نوعان) اختلف في	قَوْلُهُ (الْعِلْمُ نَوْعَانِ) أُخْتُلِفَ فِي
1	اختلف في تفسير (العلم نوعان)	أُخْتُلِفَ فِي تَفْسِيرِ (الْعِلْمُ نَوْعَانِ)
2	العلم نوعان) اختلف في تفسير العلم	...الْعِلْمُ نَوْعَانِ) أُخْتُلِفَ فِي تَفْسِيرِ
3	نوعان) اختلف في تفسير العلم فقيل	...نَوْعَانِ) أُخْتُلِفَ فِي تَفْسِيرِ الْعِلْمِ
4	اختلف في تفسير العلم فقيل لا (	أُخْتُلِفَ فِي تَفْسِيرِ الْعِلْمِ فَقِيلَ لَا (
5	اختلف في تفسير العلم فقيل لا يمكن	...أُخْتُلِفَ فِي تَفْسِيرِ الْعِلْمِ فَقِيلَ لَا
6	في تفسير العلم فقيل لا يمكن تعريفه	...فِي تَفْسِيرِ الْعِلْمِ فَقِيلَ لَا يُمْكِنُ ت

Gambar 2.4 Dataset *Windowing*

Kata-kata yang ada pada sisi kiri merupakan bentuk asal berupa arab gundul. Sementara itu, yang terdapat pada sisi kanan merupakan output yang sesuai berupa arab yang lengkap dengan harakat.

2.4. Tokenization

Tokenisasi merupakan pembagian teks input, yang ke komputer hanyalah satu string panjang karakter, menjadi subunit, yang disebut token. Token ini kemudian dimasukkan ke dalam langkah-langkah pemrosesan bahasa alami berikutnya seperti analisis morfologis, penandaan *wordclass* dan penguraian (12).

2.5. Sequences

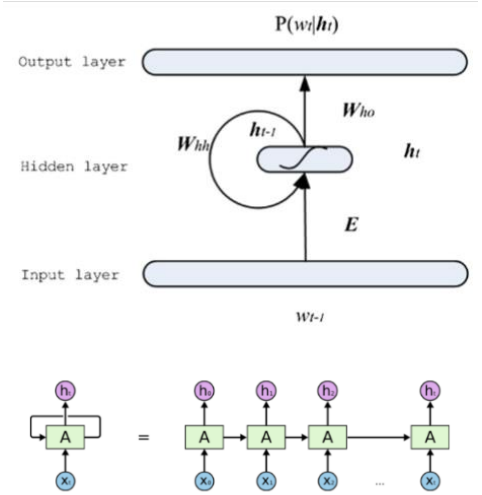
Teknik *sequences* adalah salah satu proses di dalam sebuah metode Natural Language *Processing*. *Sequences* sebagai *wordvectors* digunakan pada pembersihan data secara berulang, yang bertujuan untuk pengkodean sebuah informasi kontekstual dan menghasilkan urutan vektor baru sebagai output (13).

2.6. Padding

Pada data pelatihan, kata-kata di setiap barisnya harus memiliki panjang yang sama melalui padding. Fungsi yang digunakan untuk melakukan padding hadir di Keras sebagai *pad_sequences* (14). Berikut komposisi padding yang digunakan dalam penelitian ini, antara lain: *x_seq, maxlen = 7,* dan *padding = "post".*

2.7. RNN Model

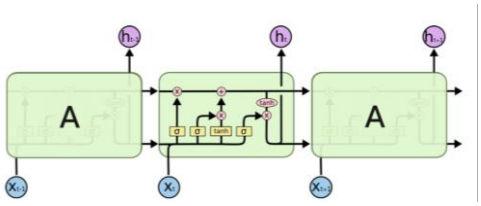
Jaringan saraf berulang Elman (RNN) ditunjukkan pada Gambar 1. RNN terkenal dengan memory panjang dan banyak digunakan untuk pemodelan sistem dinamis dan prediksi penggunaan. LetVdenotes kosakata (15)(16) Pelatihan RNN pada penelitian ini didasarkan pada x dan y. Parameter termasuk *tuple* bentuk input, panjang urutan output, jumlah dataset corpus tanpa kharakat yang unik dan jumlah dataset corpus berkharakat yang unik.



Gambar 2.5 RNN Language Model

2.8. LSTM Model

Jaringan LSTM dibuat dengan satu tujuan dalam pikiran untuk memecahkan masalah ketergantungan jangka panjang yang dimiliki RNN tradisional. Ini adalah sifat *inheren* untuk mengingat hal-hal dari waktu yang lama.



Gambar 2.6 LSTM Model

3. HASIL PENELITIAN DAN PEMBAHASAN

3.1 Aristekktur LSTM

Informasi tentang arsitektur LSTM dan parameter yang digunakan dalam penelitian ini ditunjukkan pada Gambar 3.1.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 7, 16)             422416

 bidirectional (Bidirectiona  (None, 7, 128)           41472
 l)

 bidirectional_1 (Bidirectio  (None, 7, 256)           263168
 nal)

 bidirectional_2 (Bidirectio  (None, 7, 128)           164352
 nal)

 time_distributed (TimeDistr  (None, 7, 64)            8256
 ibuted)

 time_distributed_1 (TimeDis  (None, 7, 34364)         2233660
 tributed)

=================================================================
Total params: 3,133,324
Trainable params: 3,133,324
```

Gambar 3.1 Arsitektur Model Lstm

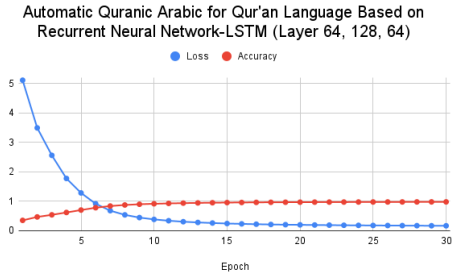
Embedding adalah representasi vektor dari kata yang dekat dengan kata-kata serupa dalam ruang n dimensional, di mana n mewakili ukuran *vektor embedding*. Alih-alih mengubah kata-kata menjadi id, penyematan kata digunakan. Input dibentuk kembali sebelum melatih jaringan saraf. Panjang indeks yang dilewati meningkat 1 untuk menghindari kesalahan indeks.

```
history = classifier.fit(x, y, validation_split=0.2, epochs=30, batch_size=10, callbacks=callbacks,
10192/10192 [==============================] - ETA: 0s - loss: 0.1746 - accuracy: 0.9778
Epoch 27: val_accuracy did not improve from 0.60529
10192/10192 [==============================] - 1591s 156ms/step - loss: 0.1746 - accuracy: 0.9778
curacy: 0.6051
Epoch 28/30
10192/10192 [==============================] - ETA: 0s - loss: 0.1724 - accuracy: 0.9787
Epoch 28: val_accuracy improved from 0.60529 to 0.60727, saving model to weights-improvement-28-0.
10192/10192 [==============================] - 1613s 158ms/step - loss: 0.1724 - accuracy: 0.9787
curacy: 0.6073
Epoch 29/30
10192/10192 [==============================] - ETA: 0s - loss: 0.1699 - accuracy: 0.9795
Epoch 29: val_accuracy did not improve from 0.60727
10192/10192 [==============================] - 1633s 160ms/step - loss: 0.1699 - accuracy: 0.9795
curacy: 0.6057
Epoch 30/30
10192/10192 [==============================] - ETA: 0s - loss: 0.1686 - accuracy: 0.9800
Epoch 30: val_accuracy did not improve from 0.60727
10192/10192 [==============================] - 1655s 162ms/step - loss: 0.1686 - accuracy: 0.9800
curacy: 0.6067
```

Gambar 3.2 Training Lstm Model

Pada Gambar 3.2 kami merancang arsitektur menggunakan *embedding* dengan jumlah kata unik ditambah satu. Sementara itu, layer yang digunakan berupa *Biderectional* dan LSTM, dengan ukuran 64, 128, 64. Pelatihan yang dilakukan terdiri dari pembagian data (*batch size*) dan *epoch.* Di mana masing-masing sebanyak 30. Informasi penjelasan *training* LSTM.

Selanjutnya pada tahapan training LSTM Model dengan performansi matrik yang dihasilkan berupa akurasi dan loss yang ditunjukkan pada Gambar 3.3



Gambar 3.3 Lstm Model Layer 64, 128, 64

Pada Gambar 3.3 menunjukkan hasil akurasi model LSTM yang baik. Dimana setiap epochnya memberikan hasil akurasi yang meningkat dan nilai loss yang semakin turun. Hasil terbaik dari keseluruhan epoch pada pelatihan yaitu dicapai pada epoch ke-30 dengan akurasi sebesar 98% dan loss 10%.

Data pelatihan sebanyak 891.800 dibagi 7 sehingga menjadi sebanyak 127.400 pada setiap bagiannya. Pelatihan yang dilakukan melalui satu kali perulangan bagi keseluruhan data. Dengan akurasi yang didapatkan yaitu 98%. Berikut hasil testing dengan mengambil kalimat pada urutan ke-1000

```
In [12]: bald_arabic = korpus_x_[1000]
         bald_arabic
Out[12]: 'صلى الله عليه وسلم بل يكون تكذيبا'

In [15]: completed_arabic
Out[15]: 'صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ بَلْ يَكُونُ تَكْذِيبًا'
```

Gambar 3.5 Testing Program

KESIMPULAN

Penelitian ini memberikan suatu inovasi baru berupa pengimplementasian algoritma Recurrent Neural Network (RNN) untuk menormalisasikan kembali aksara arab tanpa kharakat menjadi tulisan arab yang lengkap dengan kharakat. Model kami adalah model pertama di dunia yang mencoba memulihkan kata-kata Bahasa Arab yang sepenuhnya tanpa harakat dengan kalimat yang sepenuhnya berharokat menggunakan dataset training berbasis al-Qur'an dengan pencapaian akurasi 98% dengan komponen arsitektur LSTM model yang terdiri dari layer 64, 128, 64 *optimizer* "adam", *embedding* jumlah kata unik ditambah 1, dan *activation* "relu".

SARAN

Berdasarkan penelitian yang telah kami lakukan, terdapat beberapa saran pengembangan untuk penelitian selanjutnya. Pelatihan model dapat dikembangkan dengan meningkatkan akurasi dengan cara memperbanyak jumlah epoch. Sementara itu, untuk membuat tampilan lebih menarik, sangat disarankan jika dapat diimplementasikan berbasis aplikasi ataupun website agar pengguna dapat langsung menggunakannya semudah-mudahnya tanpa perlu mengoperasikan banyak tombol atau klik.

UCAPAN TERIMA KASIH

Ungkapan terima kasih kepada Prodi Teknik Informatika, Institut Teknologi Telkom Purwokerto atas dukungannya dalam menyelesaikan penelitian ini.

DAFTAR PUSTAKA

1. A. Pera Aprizal, "Urgensi Pembelajaran Bahasa Arab dalam Pendidikan Islam," *J. Pendidik. Guru*, vol. 2, no. 2, pp. 39–56, 2021, doi: 10.47783/jurpendigu.v2i2.232.
2. S. Wahyuni and R. Ibrahim, "Pemaknaan Jawa Pegon Dalam Memahami Kitab Kuning Di Pesantren," *Manarul Qur'an J. Ilm. Stud. Islam*, vol. 17, no. 1, pp. 4–21, 2017, doi: 10.32699/mq.v17i1.920.
3. N. Sa'adah, "Problematika Pembelajaran Nahwu Bagi Tingkat Pemula Menggunakan Arab Pegon," *Lisanan Arab. J. Pendidik. Bhs. Arab*, vol. 3, no. 01, pp. 15–32, 2019, doi: 10.32699/liar.v3i01.995.
4. N. Shefia, M. T. Z. Zamhuri, and F. N. Afifah, "Pemanfaatan Huruf Pegon Dalam Mempermudah Pembelajaran Nahwu," *Semnasbama*, vol. 5, pp. 189–201, 2021.
5. A. F. Rifa'i, "Analisis dan Implementasi Aplikasi Penerjemahan dan Penambah Harakat Kitab Klasik/Kitab Kuning," *Kaunia*, vol. IX, no. 2, pp. 85–95, 2013.
6. M. Alif, "Bahasa Arab Dan Problematika Transliterasi," pp. 1–10, 2020.
7. M. Musadad, A. A. J. I. Jaeni, A. Baharuddin, and A. F. Sjadzili, "Jurnal SUHUF Vol. 08 No. 01 2015," vol. 08, no. 01, 2015.
8. E. Roza, "Aksara Arab-Melayu di Nusantara dan Sumbangsihnya dalam Pengembangan Khazanah Intelektual," *Tsaqafah*, vol. 13, no. 1, p. 177, 2017, doi: 10.21111/tsaqafah.v13i1.982.
9. F. N. Rahman, "Konsep Tadarruj Dalam Internalisasi Al-Qur'an Studi Analisis Tafsir Kronologis Muhammad 'Abid Al-Jabiri Dan Teori Tahap Perkembangan Kognitif Jean Piaget," *Tesis S2 Ilmu Al Quran dan Tafsir*, 2019, [Online]. Available: http://repository.iiq.ac.id//handle/123456789/335.
10. Rodliyah Zaenuddin, "Pembelajaran Nahwu dan Sharaf dan Implikasinya Terhadap Membaca dan Memahami Literatur Bahasa Arab Kontemporer pada Santri Pesantren Majlis Tarbiyatul Mubtadiin Desa Kempek Kecamatan Gempol Kabupaten

Cirebon," *Holistik*, vol. 13, no. June, pp. 95–120, 2012.
11. Uthmani, "Quranic Arabic Corpus," *morphology, version 0.4*. Tanzil Quran Text, 2011, [Online]. Available: http://corpus.quran.com.
12. G. Grefenstette, "Tokenization," in *Syntactic Wordclass Tagging*, Vol 9. Spr., Springer, Dordrecht, 1999, p. 118.
13. M. Gardner *et al.*, "AllenNLP: A Deep Semantic Natural Language Processing Platform," 2019, doi: 10.18653/v1/w18-2501.
14. M. V. S. Rishita, M. A. Raju, and T. A. Harris, "Machine translation using natural language processing," *MATEC Web Conf.*, vol. 277, p. 02004, 2019, doi: 10.1051/matecconf/201927702004.
15. J. Ilmiah, M. Universitas, and S. Vol, "Penggunaan Aplikasi Sistem Jaringan Syaraf Tiruan Berulang Elman Untuk Prediksi Pergerakan Harga Saham Julian Talahatu, Njoto Benarkah dan Jimmy," vol. 4, no. 2, pp. 1–12, 2015.
16. M. Tomas, "Recurrent neural network based language model ´ s Mikolov Introduction Comparison and model combination Future work," *Elev. Annu. Conf. Int. speech Commun. Assoc.*, no. September, pp. 1–24, 2010, [Online]. Available: http://www.fit.vutbr.cz/research/groups/speech/servite/2010/rnnlm_mikolov.pdf.