Journal of Technology and Informatics (JoTI)
Vol. 3, No. 2, Bulan April Tahun 2022
E-ISSN: 2686-6102
P-ISSN: 2712-4842 DOI: 10.37802/joti.v3i2.224

Klasifikasi Wazan pada Kata-Kata Al Qur'an Menggunakan *Natural Language Processing*

Ira Puspasari[1], **Pranoto Hidaya Rusmin**[2]
[1,2] Teknik Elektro Institut Teknologi Bandung
Email: 33221050@ std.stei.itb.ac.id, pranoto.hr@itb.ac.id

Abstrak: Pengolahan bahasa Arab merupakan pengembangan teknik yang dapat digunakan untuk menganalisis bahasa Arab dalam konteks tertulis dan lisan. *Natural Language Processing* (NLP) memberikan kontribusi terhadap banyak sistem yang dikembangkan. Saat ini NLP telah dikembangkan dengan menggunakan teknik *Machine Learning* (ML). Algoritma ML banyak digunakan dalam NLP karena akurasinya yang tinggi. Penelitian ini membahas review penelitian pada kajian morfologi dalam Al Qur'an serta hubungannya dengan penerapan bidang komputasi sekarang, Natural Language Processing (NLP), klasifikasi wazan menggunakan NLP dengan beberapa tahapannya, termasuk pre-processing dan ekstraksi fitur. Penelitian ini menguji pola pemrosesan klasifikasi wazan menggunakan NLP dengan tahapan proses tokenization dan *Term Frequency Inverse Document Frequency* (TD-IDF). Hasil evaluasi model menghasilkan angka "1" untuk nilai *precision*, *recall*, *F1-score*, dan akurasi. Hal ini mengartikan bahwa program mampu mengklasifikasi secara tepat kata dalam pola wazan يَفْعُلُ dari pengujian sebanyak 30 data.

Kata Kunci: Al Qur'an, Morfologi, Wazan, NLP

Abstract: *Arabic processing is the development of technique that can be used to analyze Arabic in written and spoken contexts. Natural Language Processing (NLP) contributed to many systems being developed. Currently, NLP has been developed using Machine Learning (ML) techniques. The ML algorithm is widely used in NLP because of its high accuracy. This study provided literature review on the study of morphology in the Qur'an and its relationship with the application of the current field of computing, Natural Language Processing (NLP), wazan classification using NLP with its corresponding steps, including pre-processing and feature extraction. This study tested the processing pattern of wazan classification using NLP with the steps of tokenization process and Term Frequency Inverse Document Frequency (TD-IDF). The results of the model evaluation generated number "1" for the value of precision, recall, F1-score, and accuracy. This means the program is able to correctly classify words in the wazan pattern يَفْعُلُ from a test of 30 data.*

Keywords: *Al Qur'an, Morphology, Wazan, NLP*

PENDAHULUAN

Al-Qur'an adalah buku klasik dan bahasanya merupakan bahasa Arab tradisional yang dikenal sebagai i'rab [1]. Al-Qur'an memiliki kelebihan, yaitu merupakan bentuk korpus tertutup, yang meliputi: Pertama, memiliki pengulangan struktur yang sering dari frasa yang sama. Kedua, Al-Qur'an secara tradisional diidentifikasi dengan satu orang, wilayah tertentu, dan jangka waktu tertentu dan jumlahnya relatif terbatas. Kedua fakta ini membenarkan perlakuan terhadap Al-Qur'an sebagai korpus independen yang layak mendapatkan studi independen bahasanya secara umum dan sintaksis pada khususnya [2]. Memahami Al-Qur'an adalah tantangan besar bagi masyarakat, untuk pendidikan umum barat, pendidikan Muslim di dunia, representasi pengetahuan dan penalaran, pengetahuan ekstraksi dari teks, kebenaran, dan kolaborasi online. Memahami Al-Qur'an adalah hal yang utama dan merupakan tantangan besar untuk ilmu komputer dan kecerdasan buatan [3].

Linguistik komputasi merupakan persilangan antara linguistik dan ilmu komputer. Terapannya berfokus pada mengembangkan aplikasi praktis yang memiliki beberapa fasilitas dengan bahasa manusia. Saat ini aplikasi menggunakan penelitian linguistik komputasi yang tersedia termasuk: perangkat lunak pengenalan suara, web mesin pencari, pengolah kata (pemeriksa ejaan, tata bahasa checker), sistem terjemahan mesin (bahasa otomatis terjemahan). Ada banyak aplikasi yang lebih menarik saat ini sedang dikembangkan; informasi multibahasa, ekstraksi informasi, mesin terjemahan dan sistem yang bisa membaca surat kabar, jurnal dan majalah [4].

Aplikasi linguistik pada Qur'an, dibedakan menjadi dua: Bahasa Arab secara umum, dan Bahasa Arab Al-Quran. Bahasa Arab mendapat perhatian khusus oleh komunitas natural language processing (NLP), karena kepentingan politik dan perbedaannya dengan Bahasa Eropa. Karakteristik linguistik ini, memiliki morfologi yang kompleks, memberikan tantangan bagi peneliti NLP [5]. Seiring perkembangan waktu komputasi terbaru, memiliki kemajuan dan hal ini memungkinkan anotasi Al-Qur'an dengan akurasi tinggi [6].

Salah satu tujuan utama pemrosesan bahasa arab adalah pengambilan dokumen yang efektif. Misalnya, jika query adalah masukan melalui mesin pencari, yang relevan dokumen yang diambil harus didasarkan pada akar atau batang dari kata tersebut. Oleh karena itu, tujuan dari sebagian besar bahasa Arab penganalisis morfologi dan mesin pencari adalah untuk mengekstraksi akar dan/atau batang dari sebuah kata. Penelitian terbaru dilakukan di bidang Al-Quran Komputasi dapat diklasifikasikan sebagai: Pengambilan Informasi, Pengenalan Suara, Pengenalan Karakter Optik, Analisis Morfologi, Pemeriksaan Semantik, Pendidikan Aplikasi [7] Qur'an Corpus [8]. Al-Qur'an, memiliki gaya yang unik dan sifat alegoris, hal ini memerlukan perhatian khusus dalam hal masalah pencarian dan pengambilan informasi. Teknik pencarian kata kunci tidak mampu mengambil ayat-ayat semantik yang relevan [9].

Tokenisasi dalam bahasa Arab menghadirkan masalah karena kompleksitas morfologi bahasa Arab. Token biasanya didefinisikan sebagai urutan satu atau lebih. Huruf-huruf didahului dan diikuti oleh tanda. Definisi ini bekerja dengan baik untuk nonglutinasi bahasa seperti bahasa Inggris. Tokenisasi Teks-teks Arab merupakan pekerjaan yang tidak mudah [10]. Hasil dari tokenisasi memberikan pengaruh yang positif signifikan pada Named Entity Recognition (NER). Nilai F1 mengalami peningkatan saat ukuran isi -1/+1 pada puncak tokenisasi [11]. Penelitian lain tentang tiga bahasa Inggris, Arab dan Urdu, peneliti mendeskripsikan sebuah metode yang secara otomatis mengekstrak kondisi khusus pada tata bahasa lokal, dengan membandingkan perilaku token tunggal dan majemuk pada bahasa secara umum untuk menentukan token berperilaku seperti istilah atau bukan [12].

Berbagai varian dan kajian morfologi dalam Al-Qur'an merupakan keragaman pola konstruksi. Penelusuran tentang wazan-wazan dalam bahasa arab diperlukan untuk memahami proses pembentukan kata serta penggunaannya. Secara substansi dari kajian ilmu linguistik terfokus kepada empat standar substansi bahasa yaitu: Standar formasi bunyi (Standar bunyi yang mengkaji tentang eksistensi bunyi), Standar morfologi disebut ilmu bentukan kata yang mengkaji derivasi kata dan unit-unit sharafnya, termasuk wazan didalamnya, Standar syntax-grammar, dan standar semantik atau ilmu al-maany [13].

Bentrcia, pada penelitiannya tentang pendekatan untuk mengekstrak hubungan semantik dari Corpus Arab Quran, ditulis dalam aksara Arab dan menambah konstruksi otomatis ontologi Al-Quran. Penelitian ini memiliki fokus pada semantik hubungan yang dihasilkan dari pola konjungtive "DAN" [14]. Penelitian tentang metode menggunakan kategori teks untuk mengklasifikasikan kategori dengan interelasi antara berbagai sumber. Beberapa interelasi disimulasikan dengan menggunakan kombinasi sumber dataset yang berbeda dengan membandingkan Quran dan Hadis. Ketiga kategori tersebut: Haji, Sholat, Zakat yang diklasifikasikan menggunakan metode (Naïve Bayes (NB), Support Vector Machine (SVM), dan K-Nearest Neighbour (KNN)) dengan kondisi pembobotan, *Frequency– Inverse Document Frequency* (TF-IDF) [15]. Penelitian ini berfokus pada tools dan sumber korpus untuk analisis dan pemodelan Standar Arab Modern, data yang dihasilkan adalah Qur'an arab sebagai data set untuk Artificial Intelligence and Machine Learning research [16].

Cherif et al, melakukan penelitian tentang text mining menggunakan data yang cukup besar dengan membangun rule sintaks berbasis stemmer untuk meningkatkan efektivitas dalam pencarian kata pada bahasa arab [17]. Penelitian [18] tentang analisis novel yang berjudul "Qātilu Hamzah" untuk mencari wazan atau pola dan jenis-jenis jama' takṡīr berdasarkan wazan, jenis jama' takṡīr , dan juga tanda-tanda gramatikal jama' takṡīr, dengan sistem manual dan analisis dihasilkan dari 61 data, peneliti mengklasifikasikan sesuai dengan pola derivasinya serta fungsi dan kedudukan terhadap data yang ditemukan sesuai dengan pola wazn (bentuk) jama' takṡīr. Penelitian tentang "kata penting" telah dilakukan dengan Inferensi menggunakan Bidirectional LSTM model dan Inner-Attention [19]. Klasifikasi text dengan berbagai kategori telah dilakukan oleh [20] menggunakan Support vector machine, Naïve Bayes and Neural Network, hasil presisi dengan menggunakan 600 input layer sebesar 0.778, 0.754, and 0.717.

Proses anotasi melibatkan segmentasi morfologi proofreading, bagian dari speech tag dan fitur infleksi, serta menggunakan dependensi tata bahasa. Penelitian tentang penganalisis morfologi berbasis aturan, yang digunakan untuk penandaan offline awal Quranic Arabic, menghasilkan anotasi otomatis dengan skor akurasi F-measure sebesar 77% [21]. Adeleke et al. [22], mereview penelitian tentang klasifikasi versi ayat menggunakan empat metode (Decision tree, kNN, SVM, NB) untuk memperoleh keefektifan, sejumlah 1.227 ayat digunakan untuk data training dari 6.236 ayat, hasilnya Naive Bayes (NB) memiliki akurasi tertinggi (99.9099%) dan error (0.0901%). Penelitian [23] tentang isi representasi Al Quran pada NLP yang dibagi menjadi dua, yaitu: lokal dan distribusi, hal ini digunakan untuk Algoritma "machine learning" and "deep learning" yang diperlukan pada NLP.

Memahami makna wazan pada Al Qur'an akan memberikan pengetahuan tentang perbedaan makna dari kata yang ditulis menggunakan susunan huruf yang sama. Wazan berjumlah 29 dengan berbagai bentuk derivasinya dengan bentuk kata yang terdiri dari tiga, empat, lima, dan enam huruf. Wazan merupakan pola mendasar yang ada di dalam ilmu bahasa Arab, pola tersebut terdiri dari tiga buah huruf asli: 'ain fi'il (ع), fa' fi'il (ف), dan lam fi'il (ل). Jika digabung, membentuk sebuah kata yaitu فَعَلَ yang artinya mengerjakan. Pembahasan wazan yang meluas membuat peneliti mengembangkan NLP sebagai kajian morfologi untuk proses klasifikasi wazan. Penelitian ini merupakan penelitian awal dengan mengambil hanya satu jenis wazan Yaf'ulu, dengan harapan bisa diterapkan pada jenis wazan yang lain karena wazan ini termasuk pola dasar yang paling sederhana sehingga bisa diterapkan

dengan membuat aplikasi yang memilii tingkat kompleksitas rendah. Beberapa penelitian dan latar belakang dibahas pada Bagian I Pendahuluan, bagian II membahas tinjauan pustaka yang berisi tentang penjelasan Al Qur'an, pola pada kata-kata Al Qur'an, *Natural Language Processing* (NLP) dan dataset. Bagian III merupakan metode penelitian klasifikasi wazan menggunakan NLP dengan beberapa tahapannya, Bagian IV merupakan pengujian serta pembahasan penelitian sedangkan Bagian V adalah kesimpulan dan saran.

Al Qur'an

Al Quran terbagi menjadi 30 juz, 114 surat, yang dibagi menjadi dua versi makiyah dan madinah, memiliki total lebih dari 78.000 kata. Tabel 1 menunjukkan kandungan secara statistik dari Al Qur'an [24].

Tabel 1. Kandungan Al Qur'an [3]

Isi	Total	Isi	Total
Halaman	604	Ayat	6236
Bagian	30	Ayat Makki	4613
Halaman per bagian	20-23	Ayat Madani	1623
Section	60	Kata	78245
Surat	114	Kata unik	14870
Surat Makki	86 (75.4%)	Kata vokal	19273
Surat Madani	28 (24.6%)	Karakter	332837

Kandungan dari Al-Qur'an diselingi oleh satu set tanda-tanda, baik dalam baris yang sama dari ayat atau di atasnya, masing-masing memiliki arti sendiri. Beberapa tanda diantaranya: tanda berhenti (tanda Wakaf), tanda sujud, tanda bab, dan tanda ayat.

Al Qur'an ditulis dari kanan ke kiri dengan bahasa Arab. Terdiri dari 28 huruf dasar (ا، ب، ت، ث، ج، ح، خ، د، ذ، ر، ز، س، ش، ص، ض، ط، ظ، ع، غ، ف، ق، ك، ل, م، ن، ه، و، ي) dan dimodifikasi (ة،ى،أ،إ،ؤ،ئ،ء).

Wazan

Penelitian tentang pencarian dan ektraksi informasi text klasik menggunakan ontology berbasis semantik [25] mengungkapkan bahwa dengan menggunakan frame work ini ontology dapat mendeskripsikan struktur dari Qur'an bagian Juz, surat dan ayat. Dokumen ini diberikan anotasi menggunakan metode NLP seperti pemberhentian kata, stemming, dan analisis morfologi. Dalam analisis morfologi terdapat kata benda, kata kerja dan pola yang disebut dengan wazan. Struktur bahasa arab mengenal Subjek (Laki-laki, perempuan, dan Jamak), serta dikategorikan berdasarkan Madhi, Mudhari dan Amr.

Pola dasar tersebut mengalami perubahan yaitu [18]:

1. Wazan فُعُوْلٌ (fu'ūlun) : Perubahan bunyi vokal, afiks vokal panjang (infiks)
2. Wazan فَعِلَاءُ (af'ilāu) : Afiks konsonan (konfiks) dan perubahan bunyi vokal panjang
3. Wazan فِعَالٌ (fi'ālun) : Perubahan bunyi vokal dan afiks vokal panjang (infiks)
4. Wazan فُعَّالٌ (fu'ālun) : Perubahan bunyi vokal, penanggalan vokal panjang, geminasi, dan afiks vokal panjang (infiks)
5. Wazan فَعَّالٌ (fa'ālun) : Perubahan bunyi vokal, penanggalan vokal panjang, geminasi, dan afiks vokal panjang (infiks)
6. Wazan فُعَلَاءُ (fu'alāu') : Perubahan bunyi vokal, afiks bunyi vokal panjang (infiks), dan afiks konsonan (sufiks).

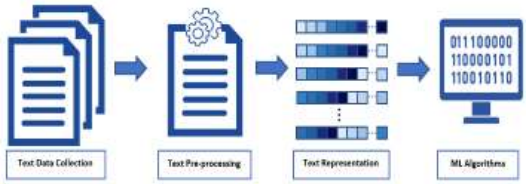
Beberapa contoh kata pada wazan يَفْعُلُ فَعَلَ terdapat pada Tabel 2.

Tabel 2. Contoh Kata pada Wazan يَفْعُلُ

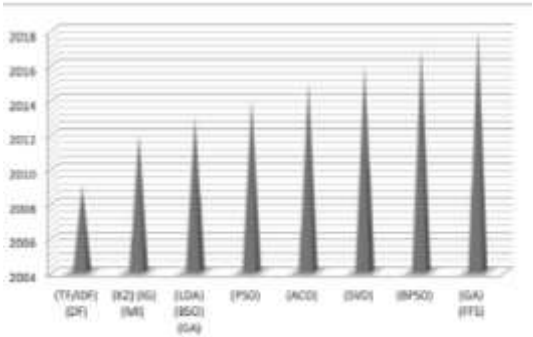
No.	Kata	Surat ke-	Ayat ke-
1	نَعْبُدُ	1	5
		2	133
		26	71
2	يَحْكُمُ	2	113
		3	47
		4	58
		68	36
		68	39
3	يَخْلُقُ	3	47
		5	17
		7	191
		16	17
		16	20

Natural Language Processing (NLP)

NLP merupakan bentuk menyeluruh dari *Artificial Intelligence* (AI).



Gambar 1. NLP-tahapan *machine learning* [23].



Gambar 2. Metode ekstraksi fitur pada Tahun 2004-2018 [26].

Sebagai simulasi pikiran manusia, teknologi ini membantu mesin untuk memahami dan menganalisis bahasa manusia. Untuk meningkatkan kinerja algoritma machine learning (ML), teknik ini digunakan untuk mewakili kumpulan data yang digunakan untuk mencapai tujuan yang diinginkan. Seperti yang ditunjukkan pada Gambar 1. Tahap analisis data pertama

terdiri dari pengumpulan data dan kemudian memprosesnya agar siap digunakan untuk representasi, yang merupakan tahapan ke penerapan algoritma pembelajaran mesin.

Data Set

Langkah pertama dalam studi Machine Learning berbasis NLP adalah pengumpulan data. Data ini adalah sampel teks yang cocok untuk bidang subjek yang bersangkutan. Seperti, Al-Nahar, Al-Jazeera, Al-Ahram, Al-hayat, dan Koran Al- Dostor, Hadith korpus, Akhbar-Alkhaleej korpus, Arabic NEWSWIRE, Quranic Arabic korpus, korpus Watan-2004, KACST Arabic korpus, BBC korpus, CCN korpus dan Open Source Arabic Corpora (OSAC), NADA korpus [26].

Klasifikasi Wazan Menggunakan Nlp

Beberapa tahapan klasifikasi wazan diantaranya adalah pre-processing pengolahan data sebelum masuk ke ekstraksi fitur yang kemudian menjadi data masukan pada Machine Learning. Paper ini mereview beberapa tahapan pre-processing, ekstraksi fitur yang dipakai pada NLP serta beberapa model *Machine Learning*.

Pre-processing

Teks *pre processing* merupakan tahapan awal pengolahan teks. Pada NLP, MADAMIRA dan RapidMiner didalamnya terdapat pemrosesan bahasa. MADAMIRA menyediakan studi tentang struktur kata-kata dan bagian dari kata-kata (akar kata-kata, preks, dan sufxes) dalam bahasa Arab. Sistem pemrosesan Bahasa Arab MADA dan AMIRA.

RapidMiner terdiri dari banyak operasi preprocessing termasuk stemming, cleaning, dan visualization, dapat dioperasikan dengan sistem operasi apa pun. Pembersihan Data, normalisasi, tokenisasi, dan stemming adalah operasi *pre processing* teks umum di sebagian besar Aplikasi NLP.

i. Pembersihan Data: terdiri dari penghapusan dan/atau koreksi catatan yang salah dari kumpulan data.
ii. Normalisasi: berfokus pada penghapusan ketidakkonsistenan variasi teks Arab.
iii. Tokenisasi: bertujuan untuk mendeteksi dan memisahkan kata-kata dengan menghilangkan komponen tambahan seperti Tanda baca, ruang putih, dan karakter unik.
iv. Stemming: digunakan untuk mengurangi berbagai bentuk kata, sehingga satu bentuk menghasilkan akar atau batang.

Feature Selection (FS)

Pengurangan dimensi atau pemilihan fitur adalah pusat dalam pengenalan pola, terutama dalam aplikasi NLP. FS bertujuan untuk meningkatkan efisiensi dan akurasi NLP dengan memilih kata-kata yang relevan. Tidak semua fitur (kata-kata dari dokumen teks) berguna untuk tahap klasifikasi karena dimensi fitur mempengaruhi kinerja klasifikasi. Untuk mengatasi tantangan tersebut, banyak fitur metode seleksi digunakan dalam penelitian NLP seperti term frequency/inverse document frequency (TF/IDF).

Selain itu, Chi-Squared statistics (X2), information gain (IG), Mutual information (MI), dan *document frequency* serta information gain juga digunakan pada seleksi fitur. Tambahan lainnya antara lain: Latent Dirichlet Allocation (LDA), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Bee Swarm Optimization (BSO), Genetic Algorithm (GA), Singular Value Decomposition (SVD) dan Fire_y Algorithm (FFS). Gambar 2. merupakan *state of the art* dari beberapa metode ekstraksi fitur pada Tahun 2004-2018.

Supervised Machine Learning Techniques

Teknik pembelajaran mesin telah berkembang selama yang terakhir dekade dan telah berguna dalam domain yang berbeda termasuk NLP, sehingga menghasilkan beberapa perangkat lunak pemrosesan bahasa yang supervised dan non supervised [26].

a. Support Vector Machines (SVMs merupakan salah satu teknik pembelajaran mesin yang diawasi. SVM telah digunakan secara efektif dalam banyak masalah. terkait dengan pengenalan pola seperti bioinformatika dan biometrik. Mengenai pemrosesan teks, SVM memiliki hasil terbaik dalam kategorisasi teks dan digunakan secara luas dalam masalah terkait NLP dalam berbagai bahasa seperti bahasa Arab untuk metode seperti prediksi keterbacaan [27].
b. Klasifikasi Naive Bayes (NB) adalah pengklasifikasi yang paling mudah dan paling banyak digunakan kedua. Merupakan teknik yang lazim untuk kategorisasi teks yang menetapkan dokumen ke dokumen terkait kategori seperti spam atau asli dan positif, negatif, atau netral [28]
c. Decision Trees telah digunakan dalam banyak masalah klasifikasi terkait NLP. Selain pengklasifikasi Naive Bayes, metode ini memberikan hasil yang sangat baik untuk deteksi spam. Decision Trees adalah teknik pembelajaran mesin yang banyak diminati karena modelnya mudah dimengerti [29].
d. K-Nearest Neighbor (k-NN) telah berhasil diterapkan pada beberapa masalah yang berkaitan NLP karena kesederhanaannya (misalnya, Ekstraksi Semantik Hubungan antara Konsep. K-NN terdiri dari pembelajaran berbasis instans, atau pembelajaran “malas”, di mana pembelajaran ditunda sampai klaasifikasi dilakukan [30].

METODE PENELITIAN

Terdapat beberapa tahapan sebelum proses klasifikasi pada penelitian ini, diantaranya adalah proses pre-processing, proses ini dilakukan untuk memilih data kata yang termasuk dalam wazan yaf'ulu dan bukan termasuk wazan yaf'ulu.

Pre-processing

Dataset pada penelitian ini dibuat dengan memilih beberapa kata pada dataset Qroots, data awal ditunjukkan pada Gambar 3.

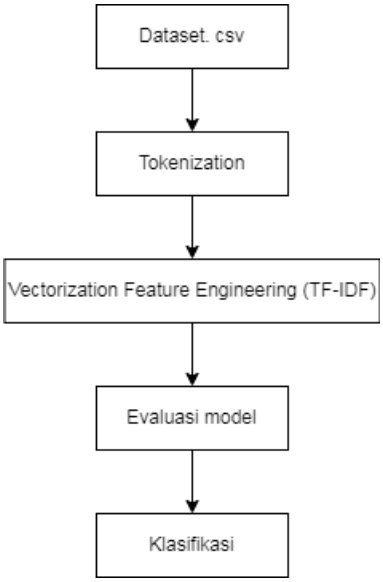
	A	B	C	D	E	F	G	H
1	nourut	posisi	surahno	ayatno	surahayat	wordno	word	wazan yaf'ulu
2	1	1	1	1	1:1	1	بِسْمِ	
3	2	1	1	1	1:1	2	ٱللَّهِ	
4	3	1	1	1	1:1	3	ٱلرَّحْمَٰنِ	
5	4	1	1	1	1:1	4	ٱلرَّحِيمِ	
6	5	2	1	2	1:2	1	ٱلْحَمْدُ	
7	6	2	1	2	1:2	2	لِلَّهِ	
8	7	2	1	2	1:2	3	رَبِّ	
9	8	2	1	2	1:2	4	ٱلْعَٰلَمِينَ	
10	9	3	1	3	1:3	1	ٱلرَّحْمَٰنِ	
11	10	3	1	3	1:3	2	ٱلرَّحِيمِ	
12	11	4	1	4	1:4	1	مَٰلِكِ	
13	12	4	1	4	1:4	2	يَوْمِ	
14	13	4	1	4	1:4	3	ٱلدِّينِ	
15	14	5	1	5	1:5	1	إِيَّاكَ	
16	15	5	1	5	1:5	2	نَعْبُدُ	يَفْعُلُ
17	16	5	1	5	1:5	3	وَإِيَّاكَ	
18	17	5	1	5	1:5	4	نَسْتَعِينُ	
19	18	6	1	6	1:6	1	ٱهْدِنَا	

Gambar 3. Pemilihan kata yang termasuk wazan.

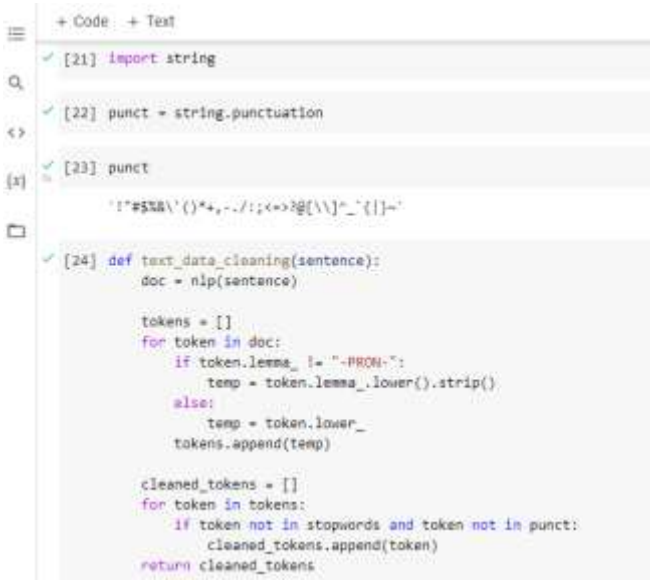
Setelah dilakukan pemilihan beberapa kata, proses selanjutnya adalah membuat dataset untuk proses training. Gambar 4 merupakan dataset yang digunakan pada penelitian ini, telah diberikan label untuk angka 1 merupakan wazan dan 0 bukan wazan. Penelitian klasifikasi wazan ini telah dibuat sebanyak total 92 data untuk pelatihan, dengan target error <0.1. Tahapan alur progam untuk klasifikasi wazan ditunjukkan pada Gambar 5. Langkah pertama adalah persiapan data seperti pada sub bab yang dibahas sebelumnya. Proses berikutnya adalah tokenisasi, pada tahap ini adalah membuat komputer memahami teks, perlu memecah kata tersebut dengan cara yang dapat dipahami mesin. Konsep ini penting pada *Natural Language Processing* (NLP). Proses tokenization pada ditunjukkan pada Gambar 6. Setelah proses tokenization selanjutnya proses *Term Frequency-Inverse Document Frequency* yang pada dasarnya memberitahukan pentingnya suatu kata dalam korpus atau dataset. TF-IDF berisi dua konsep Term Frequency (TF) dan Inverse Document Frequency (IDF). Penelitian ini kata pada korpus diubah menjadi vektor, untuk prosesnya ditunjukkan pada Gambar 7. Proses akhir setelah tercapai bentuk model yang diinginkan adalah klasifikasi dimana pada penelitian ini klasifikasi menggunakan binary text classification dengan angka "1" menunjukkan bahwa kata terssebut termasuk wazan Yaf'ulu", dan angka "0" bukan termasuk pola wazan Yaf'ulu".

Gambar 4. Labelling dataset wazan.

.



Gambar 5. Alur program klasifikasi wazan.



```
+ Code  + Text

[21] import string

[22] punct = string.punctuation

[23] punct

     '!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'

[24] def text_data_cleaning(sentence):
         doc = nlp(sentence)

         tokens = []
         for token in doc:
             if token.lemma_ != "-PRON-":
                 temp = token.lemma_.lower().strip()
             else:
                 temp = token.lower_
             tokens.append(temp)

         cleaned_tokens = []
         for token in tokens:
             if token not in stopwords and token not in punct:
                 cleaned_tokens.append(token)
         return cleaned_tokens
```

Gambar 6. Tokenization

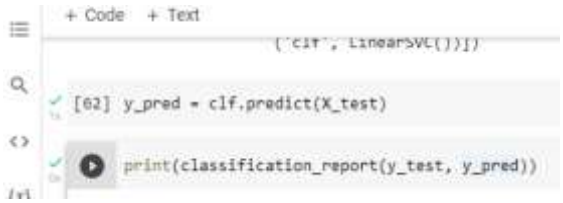
Gambar 7. Proses TD-IDF.

HASIL DAN PEMBAHASAN

Bagian ini membahas tentang hasil dari tahapan proses dari pre-processing, sampai dengan tahapan evaluasi data.

Pengujian Evaluasi Model NLP

Langkah awal adalah menguji model, sebelum dilakukan testing untuk data yang lain, Gambar 8 merupakan program untuk menampilkan hasil evaluasi dari NLP yang telah dibuat:



Gambar 8. Program menampilkan hasil evaluasi.

Hasil dari evaluasi ditunjukkan pada Gambar 9. Terdapat nilai precision, recall, F1-score.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	34
1	1.00	1.00	1.00	22
accuracy			1.00	56
macro avg	1.00	1.00	1.00	56
weighted avg	1.00	1.00	1.00	56

Gambar 9. Hasil Evaluasi.

Pengujian dilakukan pada 30 jumlah kata untuk menguji kesesuaian pengelompokan kata berdasarkan klasifikasi wazan atau bukan wazan. Terdapat beberapa batasan dalam penelitian ini antara lain: wazan dibatasi hanya 1 jenis yaitu pada bentuk يَفْعَلُ, deteksi hanya dua subjek: laki-laki dan perempuan serta posisi waktu wazan adalah mudhari. Pengujian untuk pengenalan wazan dengan hasil "1" yang berarti sesuai dengan pola wazan yaf'alu dtunjukkan pada Gambar 10. Pengujian untuk pengenalan wazan dengan hasil "0" yang berarti sesuai dengan pola wazan yaf'alu ditunjukkan pada Gambar 11. Tabel 3. hasil klasifikasi pengujian 30 kata pada data.

Gambar 10. Pengujian wazan.

Gambar 11. Pengujian bukan wazan.

Tabel 3. Hasil Pengujian Klasifikasi Kata "Wazan Yaf'ulu"

No.	Kata	Hasil Klasifikasi
1	يَخْلُقُ	1
2	يَخْلُقُكُمْ	1
3	يَسْجُدُونَ	1
4	أَصْحَٰبُ	0
5	ٱلْأُخْدُودِ	0
6	النَّارِ	0
7	نَعْبُدُ	1
8	لَيَكْتُمُونَ	1
9	يَكْتُمُونَ	1
10	يَرْزُقُ	1
11	يَشْكُرُونَ	1
12	يَرْزُقُ	1
13	يَخْلُقُ	1
14	ذَاتِ	0
15	ٱلْوَقُودِ	0
16	إِذْ	0
17	وَهُمْ	0
18	هُمْ	0
19	عَلَيْهَا	0
20	قُعُودٌ	0
21	نَعْبُدُ	1
22	يَكْتُبُونَ	1
23	يَحْكُمُ	1
24	عَلَى	0
25	وَمَا	0
26	مَا	0
27	يَفْعَلُونَ	0
28	مِنْهُمْ	0
29	لَيَحْكُمُ	1
30	يَخْلُقُ	1

Pada penelitian ini telah dibuat program klasifikasi wazan, namun klasifikasi ini hanya pada satu pola wazan yaitu: يَفْعُلُ.. Dari tahapan pre-processing data hanya dipilih 92 kata dalam Al Qur'an. Hasil evaluasi model menghasilkan angka "1" untuk nilai precision, recall, F1-score, serta akurasi. Hal ini mengartikan bahwa program ini mampu memprediksi secara tepat kata dalam pola wazan yaf'ulu ataukah bukan. Kata bukan dalam laporan ini dituliskan "0" berarti kata tersebut bisa merupakan golongan wazan yang lain seperti: فُعُوْلٌ (fu'ūlun), فِعِلَّاءٌ (af'ilāu), فِعَالٌ (fi'ālun) atau bahkan bukan bentuk-bentuk wazan. Dari hasil pengujian sebanyak 30 data telah dilakukan dan memberikan ketepatan hasil bahwa kelompok kata tersebut termasuk wazan يَفْعُلُ. atau bukan. Seperti pada kata:

يَخْلُقُ

Berdasarkan hasil pengujian klasifikasi bentuk tersebut memberikan nilai 1 yang berarti kelompok wazan yaf'ulu. Hal ini dapat dilihat dari pola huruf awal yaitu: ئ. Pola berikutnya dapat dilihat pada harokat yang sesuai dengan pola harokat يَفْعُلُ فَعَلَ yaitu:

ُ ُ ْ ُ َ

Kata yang terdeteksi bukan wazan seperti:

يَفْعَلُونَ

Berdasarkan hasil pengujian klasifikasi bentuk tersebut memberikan nilai 0 yang berarti bukan kelompok wazan yaf'ulu. Meskipun dari pola huruf awal yaitu: ئ, akan tetapi memilki harokat yang tidak sesuai yaitu:

ُ ُ َ ْ ُ

Dari 30 kata yang diklasifikasikan terdapat 15 kata yang termasuk pola wazan dan 15 kata yang termasuk bukan wazan, dengan ketepatan prediksi 100%. Hasil matrik evaluasi menunjukkan akurasi 1. Nilai akurasi menunnjukkan bahwa kata terklasifikasi secara tepat pada data uji yang diberikan. Hal ini dikarenakan data uji yang diterapkan masih sedikit jika dibandingkan dengan jumlah kata pada Al Qur'an. Data set pada penelitian ini belum terdapat di Internet karena harus membuat secara manual dengan verifikasi dari ahli sastra arab. Data penelitian ini masih sangat besar memiliki peluang untuk dikembangkan, dengan pengembangan data nantinya tidak menutup kemungkinan menurunkan tingkat akurasi sehingga kedepannya terdapat pengembangan tahapan-tahapan penelitian untuk meningkatkan akurasi.

Klasifkasi wazan فُعُوْلٌ (fu'ūlun), فِعِلَّاءٌ (af'ilāu), فِعَالٌ (fi'ālun) dan lainnya berpotensi besar dilakukan dengan menerapkan NLP, sehingga pengembangannya bisa dilakukan bahkan dengan variasi subjek tidak hanya laki-laki, namun perempuan dan jamak. Pengembangan ini dimungkinkan karena NLP merupakan proses pemahaman bahasa yang mengandung konsep penting tokenisasi, yang membuat komputer memahami teks, dengan cara yang dapat dipahami mesin dan konsep TD-IDF yang mengubah korpus menjadi vektor.

KESIMPULAN DAN SARAN

Klasifikasi wazan erat kaitannya dengan karakteristik bahasa Arab yang secara intrinsik menantang untuk pemrosesan bahasa Arab bagi para pengembang dan peneliti. Karakteristik yang paling menonjol adalah akurasi yang tinggi dan tidak adanya kapitalisasi aturan tanda baca. Beberapa tools dasar telah ditetapkan oleh peneliti NLP untuk memproses teks Arab seperti kalimat splitters, tokenizers, dan stemmers. Hasil pengujian diperoleh beberapa kesimpulan: pemrosesan klasifikasi wazan menggunakan NLP terdapat beberapa tahapan seperti: Proses tokenization dan Proses TD-IDF. Penelitian ini mampu mengklasifikasi bentuk wazan يَفْعُلُ فَعَلَ pada dataset sebanyak 92 kata pada Al Qur'an, dengan ketepatan klasifikai 100%. Klasifikasi wazan pada penelitian ini hanya untuk subjek laki-laki dan perempuan dan bukan untuk subjek jamak dengan waktu "Mudhari". Untuk pengenalan deteksi yang lebih baik dan lengkap, penelitian selanjutnya dilengkapi dengan beberapa hal: ditambahkan dataset untuk pola Yaf'ulu, kelas wazan yang lain dan objek jamak (mereka, kalian).

DAFTAR PUSTAKA

[1] B. Rahima, Z. Samir and M. Farhi, "Extracting semantic relations from the Quranic Arabic based on Arabic conjunctive patterns,*Journal of King Saud University*, vol. 30, no. Computer and Information Sciences, p. 382–390, 2018.

[2] J. Dror, D. Shaharabani, R. Talmon and S. Wintner, "Morphological Analysis of the Qur'an," *Literary and Linguistic Computing*, vol. 19, pp. 431-452., 2004.

[3] E. Atwell, K. Dukes, A. Sharaf and N. Habash, "Understanding the Quran: A new Grand Challenge for Computer Science and Artificial Intelligence," Edinburgh, 2010.

[4] S. Rahmath and K. Abdullah, "Quranic Computation A Review of research and application," in *Quranic Computation A Review of research and application RahIEEE Xplore: International Conference on Advances in Information Technology for the Holy Quran and Its Sciences*, Taibah, 2013.

[5] Y. Salem, "A Generic Framework for Arabic to English Machine Translation of Simplex Sentences Using the Role and Reference Grammar Linguistic Model and Engineering," School of Informatics at the Institute of Technology Blanchardstown, Blanchardstown, 2009.

[6] K. Dukes, "Computational Analysis of the Quran through Traditional Arabic Linguistics," 2011.

[7] A.-K. H.S, M. Al-Yahya, A. Bahanshal and I. Al-Odah, "“SemQ: A Proposed Framework for Representing Semantic Opposition in the Holy Quran using Semantic Web Technologies," in *CTIT-2009*, Dubai, 2009.

[8] D. K and T. Buckwalter, "A Dependency Treebank of the Quran using Traditional Arabic Grammar," in *INFOS 2010*, Cairo, 2010.

[9] M. Shoaib, M. Yasin, K. H. Ullah and M. M.I. Saeed, "Relational WordNet model for semantic

search in Holy Quran," in *2009 International Conference on Emerging Technologies*, Islamabad, Pakistan, 2009.

[10] A. Farghaly, "Arabic Natural Language Processing: Challenges and Solutions," *ACMTransactions on Asian Language Information Processing,* vol. 8, no. 12, 2009.

[11] B. Yassine, D. Mona and R. Paolo, "Arabic Named Entity Recognition: An Svm-Based Approach," In *First Arab International Conference and Exhibition on The Uses of White Cement*, Cairo, 2008.

[12] M. D. Rehab and Q. Islam, "Arabic Sentiment Analysis using Supervised Classification," in *Arabic Sentiment Analysis using Supervised Classification, RehabIEEE: 2014 International Conference on Future Internet of Things and Cloud*, Barcelona, 2014.

[13] M. Dr. Amrah Kasim, "Linguistiq Al Qur'an," *Jurnal Shaut Al-'Arabiyah,* vol. V, no. 1, 2017.

[14] Rahima, Z. Samir and M. Farhi," *Extracting semantic relations from the Quranic Arabic based on Arabic conjunctive patterns, Journal of King Saud University – Computer and Information Science,* vol. 30, p. 382–390, 2018.

[15] Nur, H. Nurul and H. M. Ahamed, "Text categorisation in Quran and Hadith: Overcoming the interrelation challenges using machine learning and term weighting," *Journal of King Saud University,* vol. 33, p. 658–667, 2021.

[16] Atwell, "An artificial intelligence approach to Arabic and Islamic content on the internet," IEEE: *Proceedings of NITS 3rd National Information Technology Symposium*, Leeds, 2011.

[17] W. Cherif, A. Madani and M. Kissi, "Building a syntactic rules-based stemmer to improve search effectiveness for arabic language, *IEEE: 9th International Conference on Intelligent Systems: Theories and Applications (SITA-14)* , 2014.

[18] D. Agustina, Y. Yoyo And M. T. Bin Pa, "Pola Kata Jama'taksīr Dalam Novel "Qātilu Hamzah" Karya Najib Kailani (Analisis Morfosintaksis,*A Jamiy: Jurnal Bahasa Dan Sastra Arab,* Vol. 10, No. 2, Pp. 308-325, 2021.

[19] Y. Liu, C. Sun, L. Lin and X. Wang, " Learning natural language inference using bidirectional LSTM model and inner-attention," *arXiv preprint arXiv:1605.09090.,* 2016.

[20] A. H. Mohammad, T. Alwada'n and O. Al-Momani, "Arabic text categorization using support vector machine, Naïve Bayes and neural network,*Journal on Computing (JoC),* vol. 5, no. 1, 2016.

[21] K. Dukes, E. Atwell and N. Habash, "Supervised collaboration for syntactic annotation of Quranic Arabic," *Journal of Language resources and evaluation, 47(1),* vol. 47, no. 1, pp. 33-62, 2013.

[22] A. O. Adeleke, N. A. Samsudin, A. Mustapha and N. M. Nawi, "A.,Comparative analysis of text classification algorithms for automated labelling of Quranic verses,*J. Adv. Sci. Eng. Inf. Technol,* vol. 7, no. 4, p. 1419, 2017.

[23] Z. Touati-Hamad, M. R. Laouar and I. Bendib, "Quran content representation in NLP, *Proceedings of the 10th International Conference on Information Systems and Technologies*, 2020.

[24] H. Bassam, S. Azzam and E.-H. Mahmoud," Enhancing retrieval effectiveness of diacritisized arabic passages using stemmer and thesaurus,*The 19th Midwest Artificial Intelligence and Cognitive Science Conference (MAICS2008)*, 2008.

[25] O. Ahmad, I. Hyder, R. Iqbal, M. A. A. M. Murad, S. N. M. A. and M. Mansoor, " A survey of searching and information extraction on a classical text using ontology-based semantics modeling: A case of Quran,*Life Science Journal,* vol. 10, no. 4, pp. 1370-1377, 2013.

[26] S. L. Marie-Sainte, Alalyani, A. S. N., S. Ghouzali and I. Abunadi, " Arabic natural language processing and machine learning-based systems," *IEEE Access,* vol. 7, pp. 7011-7020, 2018.

[27] H. A. R. T. Khasawneh, M. N. Al-Kabi and I. M. Alsmadi, "Sentiment analysis of Arabic social media content: A comparative study,*Proc. 8th Int. Conf. Internet Technol. Secured Trans. (ICITST)*, 2013.

[28] N. H. M., S. Elmougy, A. Ghoneim, T. Hamza, "Naive Bayes classier based Arabic document categorization, *Proc. 7th Int. Conf. Inform. Syst. (INFOS)*, 2010.

[29] J. R, T. Saleh, S. Khattab and I. Farag, "Detecting Arabic spam Web pages using content analysis," *Int. J. Rev. Comput,* vol. 6, p. 18, 2011.

[30] S. S. A, A. Q. AlHamad, M. Al-Emran and K. Shaalan, "*A survey of arabic text mining in Intelligent Natural Language Proces", Switzerland: Springer,* vol. 740, 2018.