INSYST

Journal of Intelligent System and Computation

p-ISSN:2621-9220
e-ISSN:2722-1962

Volume 3 Nomor 2, Oktober 2021

Published By **Lembaga Penelitan dan Pengabdian Masyarakat (LPPM)**
Institut Sains dan Teknologi Terpadu Surabaya (ISTTS)
formerly **Sekolah Tinggi Teknik Surabaya (STTS)**

Managed By
Departement of Informatics
Institut Sains dan Teknologi Terpadu Surabaya (ISTTS)

INSYST

Journal of Intelligent System and Computation

Volume 03 Nomor 02 Oktober 2021

INSYST

Journal of Intelligent System and Computation

Volume 03 Nomor 02 Oktober 2021

Reviewer:

INSYST

Journal of Intelligent System and Computation

Volume 03 Nomor 02 Oktober 2021

Daftar Isi

JOURNAL OF INTELLIGENT SYSTEMS AND COMPUTATION

Deteksi Validitas Berita pada Media Sosial Twitter dengan Algoritma Naive Bayes

Esther Irawati Setiawan, *Teknik Informatika, Institut Sains dan Teknologi Terpadu Surabaya,*
Sugiharto Johanes, *Teknik Informatika, Institut Sains dan Teknologi Terpadu Surabaya,*
Arya Tandy Hermawan, *Teknik Informatika, Institut Sains dan Teknologi Terpadu Surabaya,*
Yuni Yamasari, *Teknik Informatika, Universitas Negeri Surabaya.*

Abstrak*— Banyaknya berita-berita online sering menarik minat masyarakat untuk membacanya, tetapi kadang dengan terlalu banyaknya berita tersebut membuat orang susah mendapatkan informasi yang terpercaya. Berita palsu merupakan kumpulan kata atau kalimat yang mengandung informasi yang tidak benar yang berupaya untuk membohongi atau mengarahkan pembaca atau pendengarnya agar mendukung atau percaya dengan isi beritanya. Penyebar berita palsu umumnya mengetahui bahwa berita yang disebarkan tidak benar. Tujuan penelitian ini adalah mendeteksi berita palsu yang tersebar pada media sosial. Dalam mengklasifikasi berita palsu, deteksi validitas berita digunakan algoritma naïve bayes sebagai kategorisasi teks berbasis pembelajaran mesin. Penelitian ini juga membangun website yang menyediakan fitur web service, pencarian berita yang ada di Twitter, dan klasifikasi berita secara manual. User interface merupakan website berbasis PHP dimana pengguna dapat melakukan interaksi secara langsung sepeti komentar, login, atau melihat artikel-artikel yang sudah diklasifikasi. Sedangkan back-end dari website ini adalah program klasifikasi teks berbasis Python. Dari percobaan yang telah dilakukan ternyata algoritma Naïve Bayes dapat digunakan untuk mengklasifikasi berita palsu. Berdasarkan eksperimen, penggunaan metode naive bayes untuk deteksi validitas berita dengan data uji media social Twitter dapat mencapai nilai akurasi dengan persentase terbaik yaitu 92% pada data ujicoba sebesar 309 artikel.

***Kata kunci*— Berita Palsu, Klasifikasi Dokumen, Machine Learning, Naïve Bayes, Natural Language Processing.**

I. PENDAHULUAN

Berita palsu alias hoax banyak ditemui di internet. Masyarakat pun sering kali tertipu dan turut berperan dalam semakin meluasnya penyebaran berita palsu di berbagai media sosial. Berita palsu merupakan kumpulan kata atau kalimat yang mengandung informasi yang tidak benar.

Esther Irawati Setiawan, Teknik Informatika, Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Jawa Timur, Indonesia (e-mail: esther@istts.ac.id)

Sugiharto Johanes, Teknik Informatika, Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Jawa Timur, Indonesia

Arya Tandy Hermawan, Teknik Informatika, Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Jawa Timur, Indonesia (e-mail: arya@stts.edu)

Yuni Yamasari, Teknik Informatika, Universitas Negeri Surabaya (e-mail: yuniyamasari@unesa.ac.id)

Berita palsu biasanya berupaya untuk membohongi atau mengarahkan pembaca atau pendengarnya agar mendukung atau percaya dengan isi beritanya. Penyebar berita palsu umumnya mengetahui bahwa berita yang disebarkan tidak benar. Salah satu model pemberitaan palsu yang paling umum adalah menyebutkan suatu foto dari suatu kejadian sebagai kejadian lainnya dengan tujuan tertentu dari penyebarnya [1].

Berita palsu juga menjadi alat untuk mengarahkan opini public, yang dibaliknya ada tunggangan kepentingan sesorang atau komunitas tertentu. Selain itu, berita palsu yang kontroversial juga dapat dimanfaatkan penyebarnya untuk mendapatkan penghasilan dengan meningkatkan kunjungan ke situs yang dipasangi iklan. Padahal, berita palsu bisa menyesatkan dan menimbulkan kerugian bagi pihak yang menjadi korban. Dampak negatif yang ditimbulkan berupa kerugian moril maupun materiil, misalnya kehilangan reputasi, bahkan juga bisa memunculkan kebencian dari pembacanya hingga membahayakan korban.

Belakangan ini, penyebaran berita palsu banyak digunakan untuk tujuan politis. Ada pihak tertentu yang ingin menjatuhkan lawan politiknya, sehingga memanfaatkan berita palsu sebagai senjata. Pihak-pihak tersebut sengaja menggunakan media sosial untuk mempopulerkan isu negatif. Hal ini bertujuan untuk memenangkan persaingan politik sehingga terkadang antar kandidat berserta pendukung masing-masing melakukan hal-hal yang menyalahi etika demi memperoleh kemenangan.

Penyebaran berita palsu di media sosial dan aplikasi pesan telepon genggam sangat marak dan masif sehingga berdampak negatif bagi masyarakat. Efek negative yang ditimbulkan adalah rasa tidak aman, kekuatiran, kebingungan, dan kekerasan. Berita palsu dapat memperparah konflik suku, agama, ras antar golongan.

Dampak negatif berita palsu yang dipercaya oleh seseorang atau sebagian masyarakat dapat menyebabkan perpecahan dalam masyarakat, hingga pertikaian antar negara. Berita palsu dapat menyebabkan pertengkaran antar kelompok atau antar orang. Berita palsu sering mengandung hasutan dan juga ujaran kebencian yang mengganggu keharmonisan masyarakat.

Deteksi validitas berita dapat dilakukan dengan menggunakan pendekatan klasifikasi teks [2]. Metode Support Vector Machine dan C4.5 dapat digunakan untuk

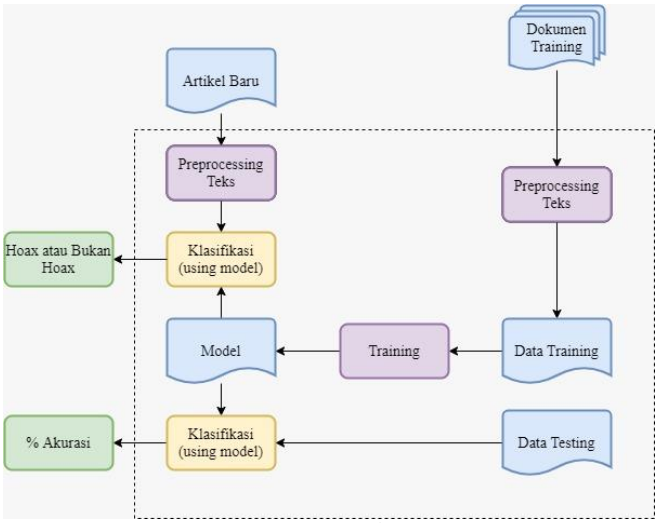
klasifikasi validitas berita [3]. Berbagai pendekatan lainnya telah digunakan untuk klasifikasi berita palsu, salah satunya adalah Stance Classification [4]. Stance Classification telah menggunakan berbagai pendekatan, yaitu Hidden Markov Model [5], Ensemble Classification [6], dan Deep Learning [7][8][9].

Sosial media adalah tempat dimana masyarakat menyampaikan pendapatnya, dan berita palsu tersebar dengan cepat. Klasifikasi dan deteksi berita palsu telah dilakukan pada Twitter [10]. Penelitian ini melakukan deteksi validitas berita pada media sosial dengan pendekatan Naïve Bayes.

Makalah ini terbagi atas lima bab utama. Bab pertama menjelaskan mengenai pendahuluan dan latar belakang. Bab kedua membahas mengenai metodologi penelitian. Bagian ketiga berisi tentang hasil dan pembahasan. Bab terakhir memaparkan kesimpulan yang diperoleh setelah melakukan penelitian.

II. METODOLOGI

Metodologi pada penelitian deteksi berita palsu ini ditampilkan pada Gambar 1. Terdapat tiga tahapan utama yaitu praproses, pelatihan dan pengujian. Pada tahap praproses, data dari media sosial diolah agar dapat digunakan untuk pelatihan maupun pengujian. Tahap pelatihan mengolah data-data cuitan dengan metode klasifikasi Naïve bayes, dan tahap pengujian bertujuan untuk melakukan klasifikasi terhadap cuitan yang ada di media sosial, apakah termasuk berita palsu atau tidak.



Gambar. 1. Arsitektur Sistem Klasifikasi Teks

Pada Gambar 1 ditampilkan arsitektur sistem pada aplikasi ini. Arsitektur sistem ini menunjukkan bagian back-end atau klasifikasi dari aplikasi karena front-end dari sistem hanya menampilkan data yang diolah di back-end.

Dokumen training yang sudah dimiliki dan dilabeli secara manual akan disiapkan sebelum klasifikasi pada tahapan preprocessing teks. Setelah semua proses tersebut selesai akan didapatkan data training yang siap dilatih. Setelah dilatih kita akan mendapatkan model training yang dapat digunakan untuk mengklasifikasi artikel baru.

Sebelum melakukan klasifikasi dengan artikel baru, akan dicoba menggunakan data training yang sudah dimiliki. Sebagian data training akan digunakan sebagai testing untuk mengetahui keakuratan dari klasifikasi. Jika akurasi yang didapatkan kurang tinggi, maka akan data training dioptimalkan kembali untuk mendapatkan akurasi yang tertinggi. Perbaikan yang dapat dilakukan adalah pemilihan data training yang lebih baik atau pemilihan penggunaan preprocessing teks yang berguna.

Jika model training sudah ada, maka dapat dilakukan pengklasifikasian menggunakan artikel baru. Proses awal dari artikel baru kurang lebih sama dengan data training. Artikel baru akan melewati proses preprocessing seperti data training. Lalu dengan menggunakan model klasifikasi yang ada akan dilakukan klasifikasi. Hasil dari klasifikasi akan muncul dengan probabilitas antara berita palsu atau bukan.

Alur kerja dalam sistem ini membutuhkan back-end untuk memberikan hasil klasifikasi kepada pengguna. User interface merupakan website berbasis PHP dimana pengguna dapat melakukan interaksi secara langsung sepeti komentar, login, atau melihat artikel-artikel yang sudah diklasifikasi. Sedangkan back-end dari website ini adalah program klasifikasi teks berbasis Python. Semua artikel yang akan diklasifikasi nantinya akan dikirim ke program Python lalu akan diterima kembali hasilnya untuk ditunjukan kepada pengguna.

Ketersediaan data training untuk deteksi validitas berita dalam Bahasa Indonesia belum ada untuk penelitian ini. Sehingga dalam penelitian ini perlu dilakukan pengumpulan data training yang sesuai agar dapat melakukan klasifikasi ke kategori yang tepat. Pengumpulan data berupa artikel berbahasa Indonesia.

A. Praproses

Preprocessing adalah pengolahan awal data text yang akan dibaca, sebelum dilanjutkan ke proses utama. Preprocessing meliputi stopword removal dan stemming. Stopword removal digunakan untuk membuang kata-kata penghubung dan stemming digunakan untuk mengembalikan kata menjadi kata dasar.

Preprocessing dilakukan terhadap data latih maupun data uji. Kemudian tahapan penghilangan stopword juga dilakukan untuk menghilangkan kata-kata yang terlalu umum pada koleksi dokumen yang dapat mengganggu performansi dan akurasi. Sedangkan tahapan stemming digunakan untuk mengembalikan kata ke bentuk dasarnya, karena varian kata-kata berimbuhan dalam Bahasa Indonesia yang cukup banyak. Berikut merupakan penjelasan lebih detail mengenai preprocessing yang dipakai dalam penelitian ini:

1. Stopwords Removal: dalam pengolahan Bahasa alami, stopword adalah kata-kata yang dihilangkan sebelum diproses algoritma klasifikasi. Daftar kata-kata ini umumnya diatur dalam stoplist. Stop word umumnya adalah kata yang yang mempunyai jumlah kemunculannya tinggi misalnya kata penghubung seperti “maka”, “sehingga”, “dan”, dan sebagainya. Terdapat berbagai pendekatan dalam penentuan stop word, dan biasanya penentuan stopword sesuai koleksi dokumen yang akan diolah. Dengan penghilangan stopword, diharapkan pengolahan klasifikasi lebih cepat dengan menghilangkan kata-kata yang tidak penting dan tidak relevan untuk diolah.
2. Stemming merupakan metode yang penting untuk praproses teks yang mengambil kata dasar sebuah kata

dengan membuang awalan, sisipan dan akhirannya. Stemming berupaya menangani berbagai varian kata yang ada pada cuitan yang sebenarnya sama kata dasarnya. Penelitian ini menggunakan stemming metode Sastrawi, sedangkan untuk penghapusan awalan, sisipan, dan akhir menggunakan algoritma dari Nazief dan Adriani.

B. *Klasifikasi dengan Naïve Bayes Classifier*

Naive bayes classifier adalah sebuah metode klasifikasi dengan teorema bayes sebagai dasarnya. Thomas Bayes, seorang ilmuan Inggris, mengusulkan metode pengklasifikasian dengan konsep probabilitas dan statistik yang memprediksi peluang di masa depan dari kejadian sebelumnya. Naive Bayes Classifier mengambil asumsi *naïve* yaitu setiap kondisi atau kejadian bersifat independen [11]. Untuk klasifikasi teks, Naïve Bayes dapat digunakan untuk penyederhanaan rumus berdasarkan Rumus 1.

$$V_{MAP} = \underset{v_j \in V}{\arg max} \frac{P(a_1, a_2 \dots a_n | v_j) P(v_j)}{P(a_1, a_2 \dots a_n)} \quad (1)$$

(a_1, a_2, …, a_n) merupakan bilangan konstan, sehingga dapat disederhanakan menjadi Rumus 2. v_j adalah probabilitas setiap kalimat terhadap sekumpulan kalimat.

$$V_{MAP} = \underset{v_j \in V}{\arg max} P(a_1, a_2 \dots a_n | v_j) P(v_j) \quad (2)$$

Naïve Bayes dapat digunakan untuk klasifikasi validitas berita pada data media sosial seperti penelitian [12] yang menggunakan Naïve Bayes untuk klasifikasi teks. Pada penelitian ini, setiap kata akan dihitung frekuensi kemunculannya dengan pendekatan Bag of Word. Kemudian dilanjutkan dengan tahapan klasifikasi menggunakan Naïve Bayes.

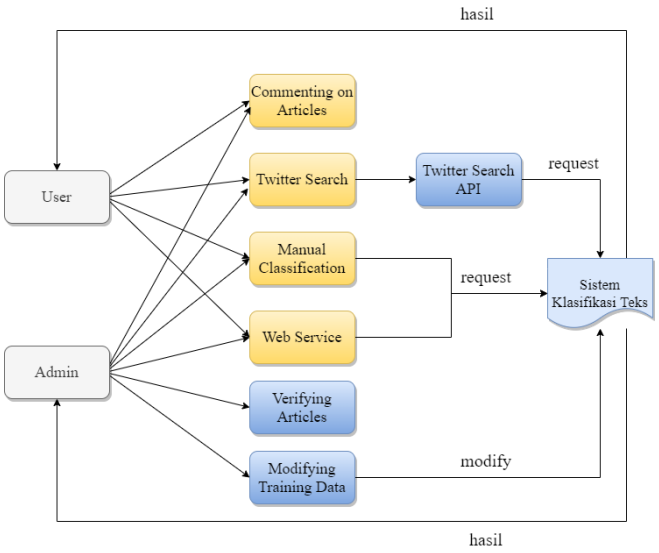
III. HASIL DAN PEMBAHASAN

A. *Aplikasi Web Deteksi Validitas Berita*

Pada bagian ini akan dijelaskan mengenai arsitektur aplikasi web yang telah dikembangkan dalam penelitian ini. Selain arsitekur umum, akan dijelaskan mengenai fitur-fitur yang ada beserta dan input-output yang dihasilkan dari penelitian ini. Gambar 2 menunjukkan artistektur umum dalam aplikasi web penelitian ini.

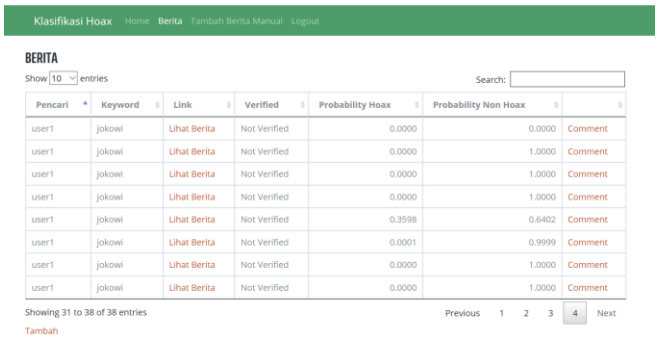
Pertama-tama pengguna melakukan login terlebih dahulu untuk dapat menggunakan website. Pengguna dapat melakukan register jika belum pernah membuat account sebelumnya. Setelah login, kemudian pengguna dapat melakukan klasifikasi teks secara manual.

Hal ini dilakukan dengan mengisi teks di tempat yang sudah disediakan. Setelah itu akan keluar hasil yang klasifikasi teks yang diharapkan. Kemudian jika pengguna ingin melakukan comment terhadap artikel-artikel yang sudah diklasifikasi, pengguna dapat pergi ke halaman berita untuk melihat semua berita yang ada. Pengguna dapat melihat artikel apa saja yang sudah diklasifikasi oleh sistem.



Gambar. 2. Arsitektur Aplikasi Web

Pengguna juga dapat melakukan search artikel-artikel yang ada di Twitter. Sistem akan melakukan request pada Twitter API ketika pengguna melakukan search query sesuai dengan keyword yang diisi. Kemudian akan ditampilkan semua hasil yang sesuai dengan keyword yang dicari beserta hasil klasifikasi yang sudah dihitung. Pengguna dapat membuka source dari berita yang sudah dicari. Output pada halaman ini akan juga menampilkan rekomendasi berita-berita yang sebenarnya. Rekomendasi berita-berita sebenarnya ini yaitu berupa dataset non-hoax yang sudah dikumpulkan.



Gambar. 3. Tampilan Halaman Admin

Pada bagian dipaparkan fitur-fitur apa saja dari web yang dibuat pada penelitian ini. Fitur-fitur yang akan ada pada penelitian ini antara lain adalah sebagai berikut:

a. Login
   Pengguna dapat melakukan pendaftaran untuk menjadi user di website ini. Walaupun begitu, pengguna lainnya tetap mampu untuk menggunakan website tanpa perlu login. Fungsi login disini hanyalah agar pengguna dapat turut serta dalam membantu berupa comment agar hasil-hasil klasifikasi yang salah dapat diperbaiki.
b. Search Article
   Pengguna dapat mencari berita yang ada di Twitter lalu secara otomatis sistem akan mengklasifikasi seluruh berita yang sesuai dengan topik yang dicari.

c. Klasifikasi Berita
Pengguna mampu memasukkan teks berita ke dalam halaman yang sudah disediakan lalu sistem akan langsung memproses teks tersebut dan mengeluarkan hasil apakah berita tersebut palsu atau bukan.

d. Comment
Disini comment berfungsi sebagai pembantu klarifikasi berita yang sudah ada. Jadi misalnya jika ada berita palsu tetapi dinilai bukan palsu oleh sistem, maka pengguna dapat melakukan klarifikasi melalui comment dan admin akan dapat melakukan tindakan lanjutan.

e. Admin Page
Halaman admin berfungsi agar admin dapat melihat dan memodifikasi dataset berita palsu dan non-palsu. Seorang admin juga dapat memoderasi comment yang ada di berita-berita, melakukan klasifikasi berita baru, dan mengganti hasil klasifikasi yang salah jika terdapat sumber yang dapat dipercaya. Tampilan halaman Admin dapat dilihat pada Gambar 3.

f. Fitur Service
Disediakan juga fitur service untuk dapat digunakan dalam aplikasi lain. Aplikasi lain dapat mengirim teks lewat URL yang sudah disediakan lalu mendapatkan kembalian berupa JSON yang berisi probabilitas teks yang sudah dikirim. Setelah user mengirim teks berita lewat url hoax/validitasberita/getclassification.php?text= akan mendapatkan kembalian berupa JSON yang berisi probabilitas hoax, non-hoax, dan kelasnya. Jika probabilitas hoax lebih tinggi berarti merupakan kelas hoax dan juga sebaliknya. User dapat menggunakan fitur ini dengan mengisi langsung teks berita lewat url dari website. Yang menjadi batasan adalah limit character url. Mengakibatkan tidak bisa mengirim teks berita yang terlalu panjang. Limit character tergantung pada browser yang digunakan oleh user.

Berikut merupakan fungsi detail beserta parameter dari web service yang disediakan:

1. URL: hoax/validitasberita/getclassification.php?text= Merupakan metode akses dari web service yang sudah disediakan dari website ini. Pengguna dapat memasukkan teks yang ingin diketahui tepat sesudah URL tersebut.
2. probabilityHoax = merupakan probabilitas hoax dari teks yang dikirimkan oleh pengguna.
3. probabilityNonHoax = merupakan probabilitas hoax dari teks yang dikirimkan oleh pengguna.
4. kelas = merupakan kelas dari teks yang dikirimkan pengguna. Terdapat 2 kelas yaitu hoax atau non-hoax. Dikatakan hoax jika probabilityHoax lebih tinggi dari probabilityNonHoax. Begitu juga sebaliknya.
5. Contoh pengembalian dan format JSON yang akan diterima:

```
{
  "probabilityHoax":"0.745338",
  "probabilityNonHoax":"0.254662",
  "kelas":"hoax"
}
```

B. Pengujian Model

Pada bagian ini akan dijelaskan uji coba dataset yang digunakan pada penelitian ini. Uji coba dataset berguna untuk mengetahui tingkat akurasi dari klasifikasi berita palsu. Terdapat beberapa scenario uji coba yang akan dilaporakan pada penelitian ini.

Pada skenario pertama, dilakukan uji coba dengan preprocessing stemming. Dataset training terdiri dari 603 artikel palsu dan 592 artikel non-palsu. Dalam uji coba dataset ini data digunakan adalah sebanyak 309 artikel, dengan rincian yaitu 152 berita palsu dan 157 non-palsu. Data uji coba sebanyak 309 ini sendiri diambil dari dataset dan tidak digunakan selama training, melainkan sebagai data data uji coba.

TABEL I
Confusion Matrix Data Uji Coba dengan Stemming

N=309	Predicted Hoax	Predicted Non-Hoax
Actual Hoax	135	17
Actual Non-Hoax	7	150
	142	167

Berdasarkan hasil uji coba pada tabel 1 didapatkan bahwa terdapat jawaban yang tidak sesuai. Namun untuk mengetahui persentase kesuksesan, dari data pada tabel confusion matrix tersebut akan dilakukan perhitungan akurasi, presisi, recall, dan F measure. Dari confusion matrix tersebut didapatkan True Positive, True Negative, False Positive, dan False Negative.

Sedangkan perhitungan akurasi, presisi dan recall, dan F1 dapat dilihat pada penjelasan selanjutnya.

True Positive (TP) = 135
True Negative (TN) = 150
False Positive (FP) = 17
False Negative (FN) = 7

$$Accuracy = \frac{150+135}{150+135+17+7} \quad (9)$$

$$= 0.922330097$$

$$Recall = \frac{150}{150+7} \quad (10)$$

$$= 0.955414013$$

$$Precision = \frac{150}{150+17} \quad (11)$$

$$= 0.898203593$$

$$F1\ Measure = \frac{2 * 0.955414013 * 0.898203593}{0.955414013 + 0.898203593}$$

$$= 0.925925926 \quad (12)$$

Karena data uji coba berasal dari data training maka sudah diketahui kelas asli dari artikel-artikel uji coba. Dari sana akan dilakukan klasifikasi terhadap data uji coba.

Setelah uji coba selesai hasil klasifikasi akan disesuaikan dengan kelas asli dari artikel-artikel tersebut. Jika kelas hasil klasifikasi uji coba sama dengan kelas asli artikel maka klasifikasi terhadap artikel tersebut dinyatakan benar. Setelah dilakukan uji coba didapatkan hasil dari True Positive, True Negative, False Positive, dan False Negative. Untuk mendapatkan akurasi dapat dilakukan pembagian dari total TP+TN dengan total TP + TN + FP + FN. Dari uji coba ini akhirnya didapatkan hasil akurasi sebesar 92%.

Skenario uji coba kedua adalah perbandingan tanpa menggunakan stemming. Seperti yang ditampikan pada tabel II, tanpa melakukan stemming diperoleh perbedaan hasil. Dapat diketahui bahwa uji coba tanpa menggunakan stemming memerlukan waktu yang lebih singkat namun akurasi yang diperoleh sebesar 84%.

TABEL II
Confusion Matrix Data Uji Coba tanpa Stemming

N=309	Predicted Hoax	Predicted Non-Hoax
Actual Hoax	114	38
Actual Non-Hoax	10	147
	124	185

Sedangkan scenario uji coba ketiga adalah penggunaan k-fold cross validation untuk mengetahui kehandalan sistem klasifikasi. Pada masing-masing fold yang diujicobakan adalah sebesar 120 artikel, yaitu 60 artikel hoax dan 60 artikel non-hoax. Tahap Preprocessing yang digunakan pada semua fold yaitu penghilangan stopwords saja. Semua fold tidak memakai preprocessing stemming.

Dataset tersebut diklasifikasikan menggunakan naïve bayes dengan pengujian 10-fold cross validation. Dapat dilihat pada Tabel bahwa nilai akurasi terbaik dihasilkan model fold-8 dengan nilai 93,33%. Dan nilai akurasi terendah adalah 85% yang dihasilkan model fold-5. Rata-rata akurasi dari 10-fold cross validation tersebut adalah sebesar 89%. Hasil 10-fold cross validation tampak pada tabel III dan grafik akurasi setiap fold ditampilkan pada Gambar 4.

TABEL III
10-fold Cross Validation

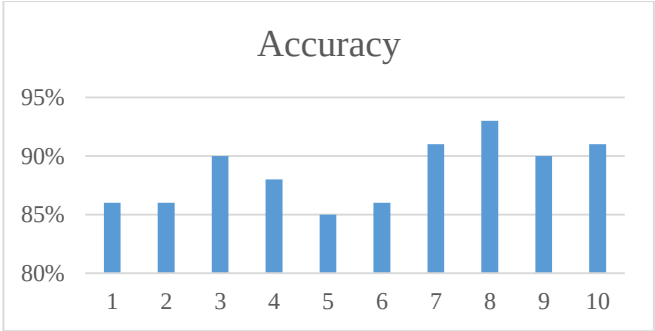
No	Accuracy
1	86,66%
2	86,66%
3	90,00%
4	88,33%
5	85,00%
6	86,66%
7	91,66%
8	93,33%
9	90,83%
10	91,07%

Selain itu, dilakukan juga uji coba performa klasifikasi dengan spesifikasi perangkat keras Intel® Core i7-4710HQ @ 2.50GHz dan RAM 8 GB. Untuk pengolahan 1.561 artikel berita dari sosial media Twitter, dibutuhkan waktu selama 2 menit sampai proses training selesai seperti yang ditampilkan pada Gambar 5.

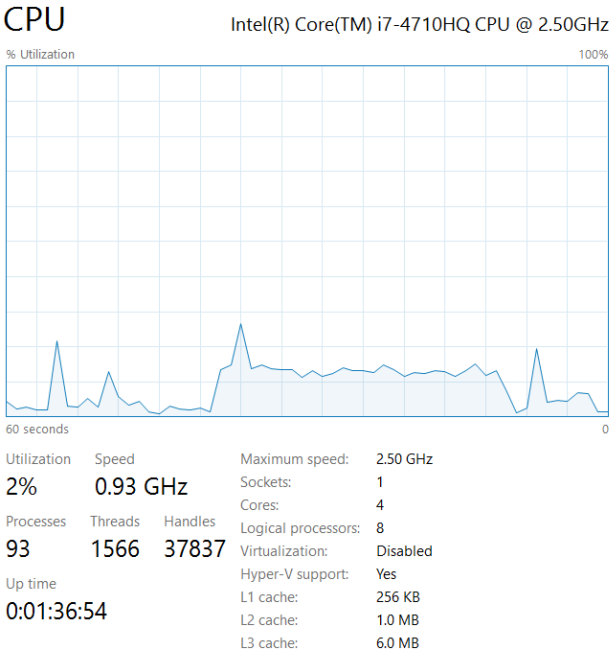
IV. KESIMPULAN

Berdasarkan proses penelitian, serta hasil dari uji coba yang telah dilakukan, terdapat beberapa kesimpulan. Kesimpulan pertama dari pengembangan penelitian ini adalah yang pertama algoritma Naïve Bayes dapat digunakan untuk mengklasifikasi berita hoax. Klasifikasi validitas berita pada media social dengan metode naive bayes menghasilkan akurasi yang baik pada data dari Twitter, dan menghasilkan nilai akurasi dengan persentase tertinggi 92% untuk data ujicoba sebesar 309 artikel.

Selain itu, preprocessing sebelum proses klasifikasi artikel berita palsu pada media social ternyata dapat mempengaruhi hasil akurasi saat uji coba. Kemudian nilai akurasi dari sistem klasifikasi diharapkan dapat meningkat dengan penambahan data training. Untuk saran penelitian selanjutnya, dapat dilakukan ensemble classification untuk deteksi validitas berita.



Gambar. 4. Grafik hasil 10-fold Cross Validation



Gambar. 5. Penggunaan CPU saat Training

DAFTAR PUSTAKA

[1] E. Lararenjana, “Mengenal Arti Hoax Atau Berita Bohong, Ketahui Jenis dan Ciri-Cirinya,” May 13, 2020. https://www.merdeka.com/jatim/mengenal-arti-hoax-atau-berita-bohong-dan-cara-tepat-menyikapinya-kln.html

[2] E. Davis, J. Adams, and S. Cohen, “Classifying articles as fake or real,” *Language and Statistics course project*, 2007.
[3] E. Rasywir and A. Purwarianti, “Eksperimen pada sistem klasifikasi berita hoax berbahasa Indonesia berbasis pembelajaran mesin,” *Jurnal Cybermatika*, vol. 3, no. 2, 2016.
[4] A. Hanselowski *et al.*, “A retrospective analysis of the fake news challenge stance detection task,” *arXiv preprint arXiv:1806.05180*, 2018.
[5] A. E. Lillie and E. R. Middelboe, “Fake news detection using stance classification: A survey,” *arXiv preprint arXiv:1907.00181*, 2019.
[6] J. Thorne, M. Chen, G. Myrianthous, J. Pu, X. Wang, and A. Vlachos, “Fake news stance detection using stacked ensemble of classifiers,” in *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, 2017, pp. 80–83.
[7] E. Lim, E. I. Setiawan, and J. Santoso, “Stance Classification Post Kesehatan di Media Sosial Dengan FastText Embedding dan Deep Learning,” *Journal of Intelligent System and Computation*, vol. 1, no. 2, pp. 65–73, 2019.
[8] E. I. Setiawan *et al.*, “Analisis Pendapat Masyarakat terhadap Berita Kesehatan Indonesia menggunakan Pemodelan Kalimat berbasis LSTM,” *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 9, no. 1, pp. 8–17, 2020.
[9] G. Rajendran, B. Chitturi, and P. Poornachandran, “Stance-in-depth deep neural approach to stance classification,” *Procedia computer science*, vol. 132, pp. 1646–1653, 2018.
[10] A. Addawood, J. Schneider, and M. Bashir, “Stance classification of twitter debates: The encryption debate as a use case,” in *Proceedings of the 8th International Conference on Social Media & Society*, 2017, pp. 1–10.
[11] A. A. Muin and others, “Metode Naive Bayes Untuk Prediksi Kelulusan (Studi Kasus: Data Mahasiswa Baru Perguruan Tinggi),” *Jurnal Ilmiah Ilmu Komputer Fakultas Ilmu Komputer Universitas Al Asyariah Mandar*, vol. 2, no. 1, pp. 22–26, 2016.
[12] D. N. Chandra, G. Indrawan, and I. N. Sukaraja, “Klasifikasi Berita Lokal Radar Malang Menggunakan Metode Naïve Bayes Dengan Fitur N-Gram,” *Jurnal Ilmiah Teknologi Informasi Asia*, vol. 10, no. 1, pp. 11–19, 2016.

Sistem Manajemen Kartu Nama dengan OCR dan Ekstraksi Informasi Otomatis

Robby Darmawan, *Departemen Informatika, Institut Sains dan Teknologi Terpadu Surabaya*,
Aris Nasuha, *Departmen Pendidikan Elektro, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia*,
Lukman Zaman, *Departemen Informatika, Institut Sains dan Teknologi Terpadu Surabaya*,
Hendrawan Armanto, *Departemen Informatika, Institut Sains dan Teknologi Terpadu Surabaya*

***Abstrak*— Sebagai pelaku bisnis, kartu nama adalah salah satu hal yang penting untuk bertukar informasi. Namun kartu nama biasanya mudah hilang atau rusak, sehingga beberapa orang biasanya menyimpan informasi dari kartu nama itu pada telepon genggam atau komputer mereka. Penelitian ini akan membuat sistem manajemen kartu nama baik individu dan juga perusahaan dengan ekstraksi informasi kartu nama otomatis untuk mempermudah pengguna perorangan ataupun perusahaan dalam melakukan penyimpanan kartu nama para kolega. Untuk mewujudkan aplikasi yang dilengkapi dengan fitur tersebut dilakukan proses pengenalan karakter pada gambar kartu nama menggunakan Tesseract OCR dan information extraction memanfaatkan klasifikasi entity dengan membangun classifier menggunakan Naive Bayes dan mengkombinasikannya dengan rule based. Hasil uji coba yang telah dilakukan mendapatkan performa 85.1% untuk pengenalan karakter dan 86% untuk pengklasifikasian entity. Dilakukan juga uji coba fungsionalitas terhadap setiap fitur pada sistem ini dengan menggunakan metode blackbox testing yang memastikan setiap aksi yang dilakukan pengguna akan menghasilkan output sesuai target yang diharapkan. Selain itu, dari hasil kuisioner yang berisikan tentang usability dari sistem ini, sebagian besar responden merasa terbantu dalam memanajemen kartu nama dengan menggunakan sistem aplikasi ini.**

***Kata Kunci*—Kartu Nama, Ekstraksi Informasi, Tesseract OCR, Klasifikasi Entity.**

I. PENDAHULUAN

Di zaman ini, dengan pesatnya usaha di segala bidang, membuat para pelaku bisnis harus mampu bersaing dengan baik untuk memasarkan produk-produk mereka. Salah satu hal yang wajib dimiliki oleh seorang pelaku bisnis untuk sukses adalah pandai menjalin relasi dengan orang lain [1].

Robby Darmawan, Departemen Informatika, FST, Institut Sains dan Teknologi Terapan Surabaya, Surabaya, Jawa Timur, Indonesia (e-mail: robby2@mhs.stts.edu)
Aris Nasuha, Departmen Pendidikan Elektro Universitas Negeri Yogyakarta, Jawa Tengah, Indonesia (e-mail: arisnasuha@uny.ac.id)
Lukman Zaman, Departemen Informatika, FST, Institut Sains dan Teknologi Terapan Surabaya, Surabaya, Jawa Timur, Indonesia (e-mail: lz@stts.edu)
Hendrawan Armanto, Departemen Informatika, FST, Institut Sains dan Teknologi Terapan Surabaya, Surabaya, Jawa Timur, Indonesia (e-mail: hendrawan@stts.edu)

Hal tersebut memaksa pelaku bisnis untuk bertemu dan berkenalan dengan banyak orang-orang baru baik sebagai rekan kerja, pelanggan, ataupun lainnya. Tentu bukanlah hal yang mudah bagi sebagian besar orang untuk bisa mengingat data kontak dan alamat seluruh orang yang mereka kenal, sehingga banyak orang menjadikan kartu nama sebagai solusi untuk hal tersebut.

Penggunaan kartu nama memiliki tujuan yang berbeda dilihat dari siapa yang memberikan maupun yang menerima kartu bisnis tersebut. Misalnya saja, untuk para pemasar, kartu nama bisnis tidak hanya digunakan untuk memberikan informasi mengenai detail informasi kontak saja tetapi juga sebagai representasi dari perusahaan Anda [2]. Biasanya sebuah perusahaan akan memiliki banyak kartu nama, baik itu pelanggan atau rekan bisnis perusahaan tersebut. Jika tidak dikelola dengan baik, biasanya data tersebut akan mudah hilang.

Semakin banyak orang tersebut berkenalan dengan orang baru, semakin banyak pula kartu nama yang akan ia miliki, di mana tentunya tidak mudah untuk melakukan penyimpanan kartu nama tersebut. Kartu nama biasanya mudah hilang dan rusak, sehingga beberapa orang biasanya menyimpan informasi dari kartu nama itu pada telepon genggam atau komputer mereka.

Mengetik informasi dari kartu nama biasanya akan membutuhkan waktu yang relatif lama terutama jika informasi yang akan dimasukkan cukup banyak. Sedangkan jika hanya difoto, informasi dari kartu nama tersebut harus diketik ulang setiap kali akan digunakan.

Berdasarkan masalah tersebut, sistem manajemen kartu nama ini akan digunakan untuk mengatasi permasalahan-permasalahan di atas. Sistem manajemen kartu nama ini akan digunakan melalui proses pengenalan informasi pada kartu nama, penyimpanan, hingga penggunaan dari kartu nama yang telah disimpan baik pada pengguna individu maupun perusahaan.

Kontribusi utama dari artikel ini adalah:

1. Membuat sistem manajemen kartu nama yang dapat mengelola data kartu nama dengan baik.
2. Membuat fitur ekstraksi informasi otomatis dari gambar kartu nama.
3. Membuat sistem manajemen kartu nama sebuah perusahaan yang dapat mengelola kartu nama dari karyawan perusahaan tersebut.

II. TINJAUAN PUSTAKA

Untuk mendukung sistem yang dibuat mengenai sistem manajemen kartu nama dengan ekstraksi informasi kartu nama otomatis dibutuhkan beberapa komponen pendukung seperti Android Studio, Laravel, Lumen, MySQL, Midtrans, OneSignal, Flask, Tesseract, NumPy, OpenCV, Scikit Learn.

A. Android Studio

Android Studio adalah Integrated Development Environment (IDE) resmi dari Android untuk pengembangan aplikasi Android [3]. IDE ini merupakan pengganti dari Eclipse Android Development Tools yang sebelumnya merupakan IDE utama untuk pengembangan aplikasi Android. Android studio pertama kali diumumkan pada tanggal 16 Mei 2013 di Google I/O conference.

B. Laravel

Laravel adalah sebuah framework yang digunakan untuk web development di PHP [4]. Laravel dibangun dengan menggunakan konsep MVC (model view controller). MVC adalah pembangunan aplikasi dengan konsep pemisahan komponen-komponen aplikasi, seperti manipulasi data, controller, dan juga user interface. Laravel dirancang untuk meningkatkan kualitas perangkat lunak dengan mengurangi biaya pengembangan awal dan biaya pemeliharaan, dan untuk meningkatkan pengalaman bekerja dengan aplikasi yang menyediakan sintaks yang ekspresif, jelas, dan menghemat waktu.

C. Lumen

Lumen merupakan salah satu micro-framework yang dikembangkan oleh Taylor Otwell, pengembang yang berada di balik framework PHP paling populer saat ini, Laravel. Lumen adalah micro framework PHP turunan dari Laravel Framework yang berorientasi pada ukuran yang kecil dan kecepatan pemrosesan halaman yang tinggi. Disebut sebagai micro framework karena ditujukan sebagai pendukung infrastruktur layanan Micro Services di server yang saat ini sedang populer digunakan. Lumen diklaim memiliki kecepatan 1900 request per detik [5], atau 100 request per detik lebih cepat dari Slim Framework versi 3 dan Silex.

D. MySQL

MySQL adalah sebuah implementasi dari sistem manajemen basis data relasional (RDBMS) yang didistribusikan secara gratis di bawah lisensi GPL (General Public License). Setiap pengguna dapat secara bebas menggunakan MySQL, namun dengan batasan perangkat lunak tersebut tidak boleh dijadikan produk turunan yang bersifat komersial. MySQL sebenarnya merupakan turunan salah satu konsep utama dalam basisdata yang telah ada sebelumnya; SQL (Structured Query Language). SQL adalah sebuah konsep pengoperasian basis data, terutama untuk pemilihan atau seleksi dan pemasukan data, yang memungkinkan pengoperasian data dikerjakan dengan mudah secara otomatis.

E. Midtrans

Midtrans merupakan payment gateway yang menyediakan layanan pemrosesan pembayaran online secara komprehensif. Midtrans menyediakan berbagai macam cara untuk transaksi pembayaran seperti menggunakan Virtual Account, GoPay, Credit Card dan Bank Transfer. Midtrans Payments menjamin proses integrasi yang mudah dan cepat di berbagai platform. Beberapa platform yang mendukung adalah iOS, Android, dan PHP.

F. OneSignal

One Signal adalah service push notification untuk website dan aplikasi mobile. OneSignal mensupport sebagian besar native dan mobile platform dengan menyediakan SDK untuk masing-masing platform, RESTful server API, dan online dashboard untuk melihat performa, statistik penggunaan maupun operasi push notification.

G. Flask

Flask adalah sebuah web framework yang ditulis dengan bahasa Python dan tergolong sebagai jenis microframework. Flask termasuk pada jenis microframework karena tidak memerlukan suatu alat atau pustaka tertentu dalam penggunaannya. Selain itu, meskipun Flask disebut sebagai microframework, bukan berarti Flask mempunyai kekurangan dalam hal fungsionalitas. Microframework disini berarti bahwa Flask bermaksud untuk membuat core dari aplikasi ini sesederhana mungkin tapi tetap dapat dengan mudah ditambahkan [6].

H. Tesseract

Tesseract adalah mesin pengenal karakter optik yang bersifat gratis. Tesseract OCR adalah proyek open-source, dimulai oleh Hewlett-Packard, kemudian Google mengambil alih pengembangan [7]. Kini, tesseract telah dianggap sebagai salah satu mesin perangkat lunak OCR bebas yang paling akurat yang tersedia.

I. NumPy

NumPy merupakan sebuah library paket dasar Python guna keperluan komputasi scientific. Selain digunakan untuk masalah-masalah scientific dan operasi aritmatika, NumPy juga dapat digunakan sebagai sebuah penampung data, seperti contoh untuk menampung dataset gambar yang digunakan pada Penelitian ini. Objek utama pada NumPy berupa sebuah array multidimensi. Objek multidimensional array ini biasa disebut sebagai sebuah ndarray (N dimensional array).

J. OpenCV

OpenCV (Open source computer vision) adalah library pemrograman yang terutama ditujukan untuk bidang computer vision. Library ini digunakan dengan tujuan untuk membantu proses pengolahan citra. Library ini akan digunakan untuk membaca gambar dan video. Selain itu, library ini juga membantu proses-proses pada data gambar, seperti resize, rotate, dan convert color. Library ini juga membantu dalam melakukan pendeteksian mata, pendeteksian wajah, juga dalam memotong atau mengambil area tertentu pada gambar.

K. Scikit Learn

Scikit-learn atau sklearn adalah modul untuk bahasa pemograman python yang dibangun diatas NumPy, SciPy, dan matplotlib, fungsinya dapat membantu melakukan processing data ataupun melakukan training data untuk kebutuhan machine-learning [8]. Ada banyak fitur yang dapat digunakan dengan sklearn ini, seperti Classification, Regression, Clustering, Dimensionality reduction, Model selection, dan Preprocessing data.

III. ANALISA SISTEM

Pada bagian ini akan dijelaskan tentang analisa sistem dari Sistem Manajemen Kartu Nama dengan Ekstraksi Informasi Kartu Nama Otomatis. Sebelum program dibuat, terdapat berbagai macam hal yang perlu disiapkan atau dilakukan terlebih dahulu. Salah satu hal yang perlu dilakukan adalah melakukan analisa terhadap permasalahan yang ditemukan. Setelah menemukan masalah dan melakukan analisa, dibuatlah kesimpulan dari hasil analisa dimana adanya kebutuhan untuk membuat sebuah program dalam menyelesaikan masalah yang sedang dihadapi.

A. Analisa Permasalahan

Untuk pelaku bisnis yang terbiasa menyimpan data kartu nama pada kontak telepon genggamnya juga terdapat masalah yakni dibutuhkan waktu yang lama untuk menginputkan seluruh informasi dari kartu nama. Selain itu, mereka juga tidak dapat menyimpan gambar dari kartu nama tersebut bersamaan dengan data kontak yang telah disimpan.

Untuk permasalahan yang dimiliki oleh perusahaan, biasanya perusahaan tidak dapat memanajemen seluruh data kartu nama baik itu kartu nama rekan bisnis perusahaan tersebut ataupun kartu nama dari pelanggan-pelanggan perusahaan tersebut. Biasanya data kartu nama itu cukup bergantung kepada karyawan-karyawan tertentu yang berhubungan langsung dengan pemilik kartu nama. Tentunya perusahaan akan sangat mudah kehilangan data kartu nama tersebut ketika karyawan tersebut keluar dan tidak bekerja di perusahaan itu lagi.

Dari analisa yang sudah dilakukan, akan dihasilkan solusi dari masalah yang bersangkutan. Analisa permasalahan adalah analisa dari masalah yang telah disebutkan pada penjelasan sebelumnya. Berikut adalah hasil analisa permasalahan yang dijabarkan dalam bentuk poin-poin:

- Pelaku bisnis sulit untuk menyimpan kartu nama secara fisik, terutama ketika jumlah kartu nama sudah sangat banyak.
- Pelaku bisnis akan membutuhkan waktu yang lama jika harus menginputkan data informasi kartu nama satu per satu.
- Pelaku bisnis tidak dapat menyimpan gambar sekaligus dengan ekstraksi informasi kartu nama pada kontak telepon genggam.
- Perusahaan seringkali kehilangan data kartu nama dari karyawan yang tidak lagi bekerja di perusahaan tersebut

B. Analisa Kebutuhan

Untuk mulai melakukan analisa, disimpulkan sebuah permasalahan paling dasar dari subbab sebelumnya yaitu mengenai manajemen kartu nama mulai dari menyimpan, menggunakan, hingga pendistribusian kartu nama ke akun perusahaan. Setelah melakukan analisa lebih lanjut, maka terdapat gagasan untuk membuat sistem aplikasi manajemen kartu nama yang memiliki fitur ekstraksi kartu nama otomatis. Dengan fitur ini, pengguna hanya perlu memindai gambar kartu nama lalu memverifikasi hasil dari ekstraksi informasi kartu nama. Dengan demikian, pengguna akan menghemat waktu dalam mengetikkan informasi kartu nama. Sedangkan untuk permasalahan pada pengguna perusahaan, maka akan dibuat sistem dengan fitur untuk memanajemen karyawan dari perusahaan tersebut. Nantinya, setiap karyawan dapat mengkontribusikan kartu nama yang mereka miliki ke perusahaan sehingga perusahaan dapat memanajemen seluruh data kartu nama dengan baik.

C. Spesifikasi Kebutuhan Dasar

Spesifikasi kebutuhan yang dijelaskan akan berupa fitur-fitur utama serta teknologi yang sangat diperlukan dalam membuat fitur utama. Fitur utama adalah fitur yang dapat mengurangi atau menyelesaikan permasalahan yang ditemukan sebelumnya. Fitur-fitur utama yang menjadi solusi akan menjadi fitur yang dibuat pada Penelitian ini. Berikut adalah fitur-fitur utama dari sistem ini:

- Ekstraksi Informasi Kartu Nama
  Fitur ekstraksi informasi kartu nama merupakan salah satu fitur utama yang akan membantu permasalahan yang ditemukan sebelumnya yakni meningkatkan efisiensi dalam penyimpanan data kartu nama. Dalam fitur ini, pengguna akan menginputkan gambar kartu nama baik melalui galeri atau mengambil gambar secara langsung dengan kamera. Nantinya gambar akan dikirim ke sebuah servis yang dibangun dengan menggunakan framework Flask.
- Manajemen Kartu Nama Perusahaan
  Fitur manajemen kartu nama perusahaan ini merupakan fitur yang akan membantu perusahaan menyelesaikan masalah dalam mengatur data kartu nama mereka. Fitur ini akan memperbolehkan perusahaan untuk mengatur karyawan-karyawan mereka ke dalam akun perusahaan mereka. Setiap karyawan bisa berkontribusi untuk menyimpan data kartu nama mereka ke akun perusahaan mereka. Nantinya akun perusahaan dapat mengelola data kartu nama tersebut dan juga membagikan kartu nama itu ke karyawan mereka yang lain, sehingga ketika ada karyawan yang keluar dari perusahaan, maka perusahaan dapat dengan mudah memindahkan atau membagikan kartu nama tersebut ke karyawan yang lain.

D. Fitur Pelengkap

Pada bagian ini akan dijelaskan mengenai penjelasan fitur-fitur yang akan membangun sistem secara lebih detail. Fitur-fitur yang dijelaskan akan menjadi fitur yang dibuat pada sistem ini sebagai pelengkap fitur utama yang telah disebutkan sebelumnya.

- Pencarian dan Pengurutan Kartu Nama
  Fitur ini akan membantu pengguna dalam melakukan pencarian kartu nama. Tidak sedikit pengguna yang memiliki data kontak atau kartu nama hingga ratusan bahkan ribuan kartu nama. Dengan fitur ini pengguna dapat memfilter data kartu nama yang akan ditampilkan.
- Membagikan Kartu Nama
  Fitur ini dapat digunakan oleh pengguna untuk membagikan kartu nama mereka kepada orang lain melalui sosial media seperti line atau whatsapp. Penerima kartu nama dapat menyimpan kartu nama yang telah dibagikan tersebut. Sedangkan untuk pengguna perusahaan, mereka dapat membagikan kartu nama antar karyawan yang mereka miliki secara massal.
- Melakukan Panggilan dan Mengirim Pesan
  Salah satu tujuan dalam menyimpan kartu nama adalah

untuk menghubungi pemilik kartu nama tersebut, baik melakukan panggilan telepon atau mengirim pesan kepada pemilik kartu nama. Oleh karena itu, fitur ini akan membantu pengguna untuk melakukan panggilan dan juga membuka halaman pembuatan pesan yang langsung memasukkan nomor tujuan pesan secara otomatis.

- Manajemen Berlangganan

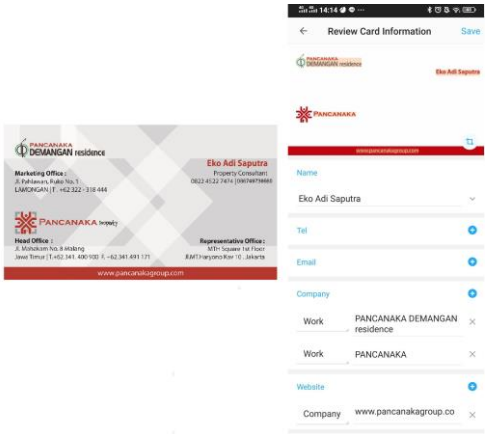
  Dalam penggunaan sistem ini, akun perusahaan akan dikenakan biaya berlangganan yang akan ditagihkan setiap bulannya. Oleh karena itu, perusahaan akan diminta untuk memasukkan data kartu kredit / kartu debit yang akan digunakan untuk melakukan pembayaran biaya langganan.
- Manajemen Karyawan

  Dalam memanajemen kartu nama, perusahaan membutuhkan karyawan-karyawan untuk berkontribusi dalam menambahkan data kartu nama. Oleh karena itu, fitur ini akan digunakan perusahaan untuk menambahkan dan menghapus karyawan. Dalam proses penambahan karyawan, admin perusahaan harus mendaftarkan nomor telepon dari karyawan dan harus menunggu konfirmasi dari karyawan tersebut. Hal ini juga bisa dilakukan dengan cara sebaliknya, pengguna individu mendaftarkan akunnya ke perusahaannya dan admin perusahaan akan mengkonfirmasinya.
- Manajemen Pengguna

  Untuk menanggulangi beberapa masalah seperti gagal melakukan pengiriman kode verifikasi atau penyalahgunaan aplikasi, maka super admin dapat memanajemen pengguna dari sistem manajemen kartu nama ini. Super admin dapat melakukan penerimaan verifikasi secara manual pada fitur ini. Selain itu, admin juga dapat memblokir pengguna baik pengguna individu maupun pengguna perusahaan yang dianggap melakukan penyalahgunaan pada sistem manajemen kartu nama ini.

E. Analisa Software Sejenis

Pada bagian ini akan dijelaskan contoh software sejenis yang sudah ada sebelumnya. Bagian ini juga akan menjelaskan perbandingan terhadap fitur-fitur yang ada pada software sejenis dengan sistem pada Penelitian ini. Ada dua software yang akan menjadi perbandingan, yang pertama adalah Tesis "Pengenalan Layout dan Ekstraksi Teks pada Citra Kartu Nama" oleh Maulidiansyah NRP 213210397 dan yang kedua adalah aplikasi Camcard.



Gambar 1. Contoh Ekstraksi Kartu Nama CamCard

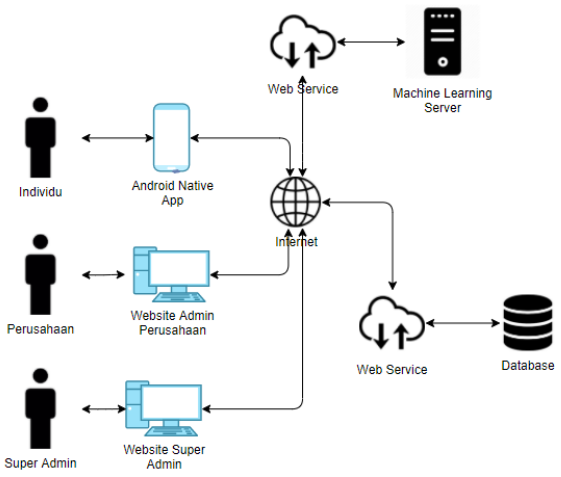
Perbedaan mendasar Penelitian ini dengan tesis "Pengenalan Layout dan Ekstraksi Teks pada Citra Kartu Nama" adalah mengenai aplikasi sistem manajemen kartu nama. Sehingga Penelitian ini, tidak hanya mengekstrak informasi kartu nama tersebut, melainkan juga dapat memanfaatkan hasil dari ekstraksi kartu nama tersebut.

Sedangkan aplikasi kedua, aplikasi Camcard, tidak jauh berbeda dengan sistem manajemen kartu nama pada Penelitian ini. Namun, dengan adanya Penelitian ini diharapkan ekstraksi kartu nama dapat lebih baik dibanding dengan aplikasi Camcard seperti pada Gambar 1. Desain

Pada bagian ini akan dijelaskan tentang desain sistem dari program yang dibuat pada sistem ini. Pada bagian ini, akan dijelaskan desain-desain yang dijelaskan berupa desain arsitektur dan desain tampilan. Desain arsitektur dibutuhkan untuk mengetahui garis besar komponen-komponen yang akan menyusun sistem secara kesuluruhan. Desain tampilan adalah bentuk atau tampilan dari program yang dibuat pada sistem ini.

F. Arsitektur Sistem

Pada bagian ini akan dijelaskan mengenai arsitektur dari sistem manajemen kartu nama ini. Pengguna sistem ini nantinya harus mengunduh dan menginstall aplikasi ini pada perangkat android untuk dapat menggunakan aplikasi ini. Proses kerja dari aplikasi ini dapat dilihat pada Gambar 2.



Gambar 2. Arsitektur Program

Dapat dilihat pada Gambar 2 bahwa dalam sistem ini, akan terdapat tiga macam aktor pengguna yakni pengguna individu, perusahaan, dan juga admin sistem ini. Pengguna individu dapat mengakses sistem ini melalui aplikasi mobile berbasis android. Sedangkan untuk pengguna perusahaan dan super admin nantinya akan mengakses sistem ini melalui website.

Dari Gambar 2 juga dapat dilihat bahwa akan ada beberapa program yang akan dibuat pada Penelitian ini agar sistem dapat berjalan dengan baik. Program tersebut adalah android mobile app, website perusahaan, website admin, web service, dan juga machine learning web service. Berikut adalah penjelasan dari masing-masing program tersebut.

- Android Mobile App

  Android mobile app pada Penelitian ini akan dibuat secara native menggunakan android studio. Android mobile app ini akan digunakan oleh pengguna individu

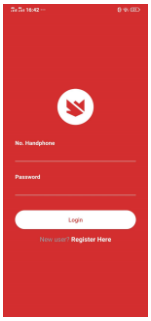
mulai dari proses scan, penyimpanan kartu nama, pencarian, dan sebagainya.

- Website Perusahaan
  Website perusahaan akan dibuat menggunakan framework Laravel. Website ini akan digunakan oleh perusahaan yang telah melakukan subscription dalam memanejemen data perusahaan tersebut. Perusahaan akan dapat menyimpan dan memanipulasi data kartu nama mereka melalui website ini.
- Website Admin
  Website admin juga akan dibuat menggunakan framework Laravel. Website ini akan digunakan oleh admin dalam mengatur seluruh user dan juga mengatur subscription dari perusahaan yang mendaftar. Penggunaan website admin ini akan dikenakan biaya untuk setiap bulannya yang akan ditagih otomatis melalui kartu kredit / debit.
- Web Service
  Web service pada sistem ini akan dibuat menggunakan Lumen. Web service ini akan digunakan untuk melayani seluruh request dari sistem untuk menyimpan dan mengambil data yang dibutuhkan.
- Entity Recognition Web Service
  Entity recognition server pada sistem ini akan dibuat menggunakan framework flask. Entity recognition web service ini digunakan untuk melayani request dalam mengekstraksi informasi kartu nama dari gambar yang dikirimkan. Dua proses utama yang dijalankan pada service ini adalah Optical Character Recognition yang dijalankan menggunakan Tesseract dan Named Entity Recognition.

G. Desain Interface

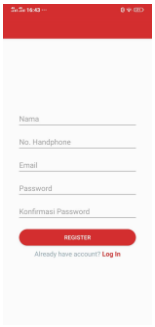
Pada bagian ini akan diberikan bentuk tampilan dari desain interface pada website dan aplikasi mobile serta penjelasan untuk masing-masing desain interface tersebut. Penjelasan desain interface akan dibagi berdasarkan fitur-fitur yang dibuat. Penjelasan desain interface fitur akan diurutkan berdasarkan kronologis penggunaan fitur secara berurutan, dan diikuti dengan fitur-fitur tambahan.

- Interface Login Register
  Pada bagian ini akan dijelaskan desain interface yang diterapkan pada fitur login dan register pada aplikasi mobile sistem manajemen kartu nama. Interface ini akan diberikan kepada pengguna ketika pengguna ingin melakukan login atau registrasi pada aplikasi.



Gambar 3. Tampilan Login

Seperti pada gambar 3 akan terdapat dua input teks, yang pertama adalah unuk nomor telepon pengguna dan kedua adalah untuk password pengguna. Bagian password menggunakan tipe input yang khusus untuk password agar tidak dapat dilihat orang lain saat pengguna mengisinya. Terdapat sebuah tombol yang dapat digunakan untuk login setelah pengguna mengisi kedua input teks. Lalu dibawahnya terdapat sebuah tombol yang dapat digunakan untuk menuju ke halaman registrasi. Berikut adalah gambar desain interface saat pengguna membuka halaman registrasi.



Gambar 4. Tampilan Register

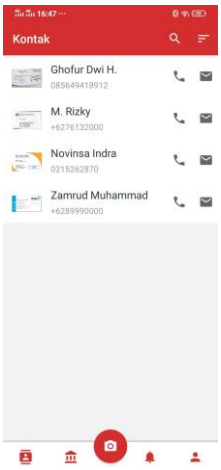
Pada gambar 4 ditampilkan bentuk dari form registrasi pengguna baru. Pada form registrasi terdapat 5 input yang harus diisi agar dapat melakukan registrasi. Input tersebut terdiri dari nama, nomor telepon, email, password, dan konfirmasi password. Pada bagian nomor telepon, pengguna wajib memasukkan nomor teleponnya dengan benar karena pengguna akan menerima kode verifikasi melalui sms yang dikirimkan ke nomor telepon tersebut.

Gambar 5. Tampilan Verifikasi OTP

Setiap pengguna yang baru mendaftarkan diri ke sistem ini, akan menerima kode verifikasi pada nomor teleponnya. Gambar 5 adalah tampilan yang digunakan oleh pengguna untuk memasukkan kode verifikasi yang mereka terima sebelumnya. Ketika pengguna gagal menerima kode verifikasinya, pengguna dapat menekan tombol kirim kembali dengan menunggu selama 2 menit. Setelah pengguna memasukkan kode verifikasi dan menekan tombol verifikasi, maka status pengguna akan menjadi aktif dan akan langsung diarahkan ke halaman daftar kontak.

- Interface Daftar Kontak
  Halaman daftar kontak adalah halaman yang digunakan untuk menampilkan seluruh kontak atau kartu nama yang dimiliki pengguna yang sedang login. Pada setiap daftar kontak yang dimiliki, akan muncul tombol dengan logo telepon dan pesan di sebelah kanan nama kontak. Dengan menekan tombol telepon, pengguna akan otomatis dilempar ke halaman dial call dengan nomor tujuan sesuai

dengan kontak yang dipilih. Begitu pula dengan tombol dengan logo pesan, jika ditekan pengguna akan diarahkan ke halaman pembuatan pesan dengan nomor tujuan sesuai dengan kontak yang dipilih. Gambar halaman daftar kontak ini dapat dilihat pada Gambar 6.

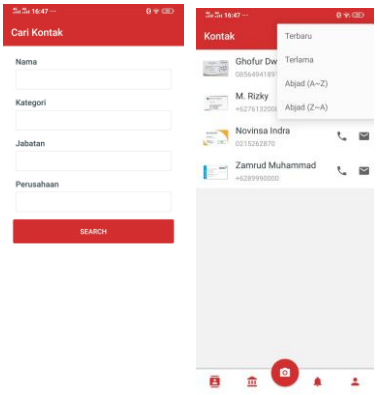


Gambar 6. Tampilan Daftar Kontak

Pada Gambar 6 juga dapat dilihat akan terdapat 5 tombol navigasi di bagian bawah layar. Tombol navigasi ini secara urut adalah tombol navigasi halaman daftar kontak, tombol navigasi halaman daftar perusahaan, tombol navigasi menambah kontak, tombol navigasi halaman notifikasi, dan yang terakhir adalah tombol navigasi halaman profil. Kelima tombol navigasi tersebut akan selalu tampil di halaman-halaman utama pada setiap menunya, sehingga akan memudahkan pengguna untuk berpindah halaman dari halaman yang satu ke halaman yang lainnya.

- Interface Pencarian Kontak

  Untuk mempermudah pengguna menampilkan data kontak yang mereka inginkan maka terdapat halaman untuk menginputkan filter sesuai dengan atribut yang diinginkan. Selain itu, pengguna juga dapat mengurutkan tampilan daftar kontak berdasarkan abjad nama atau waktu dari kartu nama tersebut disimpan. Gambar dari tampilan pencarian kontak dapat dilihat pada Gambar 7.



Gambar 7. Tampilan Pencarian Kontak

Pada Gambar 7 bagian kiri dapat dilihat bahwa terdapat 4 macam atribut yang dapat digunakan pengguna untuk mencari kartu nama yang diinginkan. Atribut itu adalah nama, kategori, jabatan, dan juga nama perusahaan. Ketika pengguna telah memasukkan data atribut yang akan dicari dan menekan tombol search, maka pengguna akan kembali dilempar ke halaman daftar kontak dengan kontak yang sesuai dengan hasil pencarian. Di halaman daftar kontak bagian kanan atas, terdapat tombol urutkan yang memiliki pilihan untuk merubah urutan tampilan berdasarkan abjad nama atau waktu kartu nama disimpan.

- Interface Daftar Perusahaan

  Halaman daftar perusahaan adalah halaman yang digunakan pengguna untuk mengatur perusahaan tempat mereka bergabung. Seperti yang dapat dilihat pada Gambar 8, pengguna dapat memasukkan email perusahaan dan menekan tombol request untuk bergabung ke sebuah perusahaan. Setelah melakukan request, maka daftar perusahaan akan bertambah di bagian bawah tampilan. Pada daftar perusahaan akan muncul nama perusahaan, email perusahaan, dan status dari pengguna. Ketika pertama kali pengguna mendaftarkan dirinya, maka status akan menjadi pending hingga perusahaan menerima permintaan bergabung pengguna.

Gambar 8. Tampilan Daftar Perusahaan

- Interface Tambah Kontak

  Halaman ini digunakan pengguna untuk menambahkan kontak atau kartu nama baru. Ketika pengguna menekan tombol navigasi tambah kontak, pengguna akan diberikan pilihan antara membuka galeri atau kamera. Tampilan pilihan tersebut dapat dilihat pada Gambar 9 bagian kiri. Setelah pengguna memilih, maka pengguna akan dilempar ke halaman sesuai pilihan pengguna. Pengguna dapat memasukkan gambar kartu nama baik dari galeri atau kamera. Setelah gambar dipilih dan dikirim ke servis, pengguna akan melihat tampilan seperti pada Gambar 9 bagian kanan. Gambar kartu nama dapat diperbesar dan digeser sesuai keinginan pengguna. Atribut dari kartu nama akan otomatis terisi seperti pada gambar, namun ketika pengguna merasa pengisian data tidak sesuai, pengguna dapat menggantinya sesuai dengan opsi pada combobox ataupun mengetiknya secara manual.

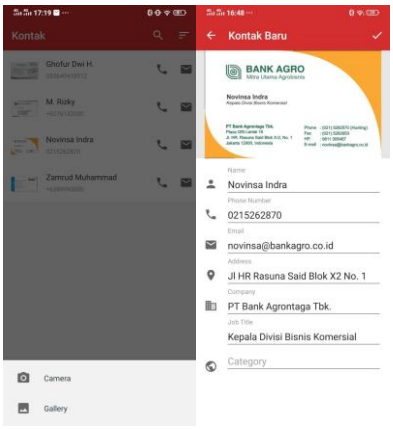
- Interface Pengaturan Profil

  Halaman pengaturan profil ini akan memiliki beberapa menu yakni untuk mengubah data pengguna, mengubah password, mengubah pengaturan notifikasi, dan informasi mengenai versi aplikasi yang terinstal. Gambar halaman profil ini dapat dilihat pada Gambar 10.

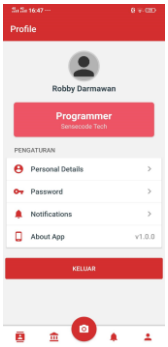
  Seperti pada Gambar 10 pada bagian atas akan muncul informasi singkat mulai dari nama, jabatan, dan juga nama perusahaan dari pengguna yang sedang login. Di bawah tampilan tersebut, akan terdapat beberapa menu seperti yang sudah disebutkan sebelumnya. Pada bagian paling

bawah terdapat tombol keluar yang dapat digunakan pengguna untuk keluar dari aplikasi.

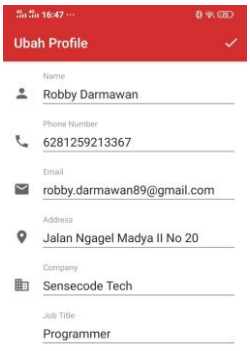
Ketika pengguna menekan tombol menu personal details, maka pengguna akan menuju halaman untuk mengubah profil. Seperti pada Gambar 11 pengguna dapat mengubah profil sesuai input yang ada. Input tersebut terdiri dari nama, nomor telepon, email, alamat, nama perusahaan, dan jabatan. Setelah memasukkan data, pengguna dapat menekan tombol simpan dengan logo centang di pojok kanan atas.



Gambar 9. Tampilan Tambah Kontak

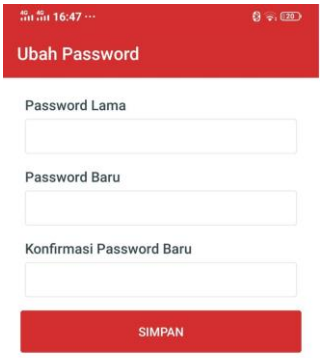


Gambar 10 Tampilan Pengaturan Profil



Gambar 11. Tampilan Mengubah Profil

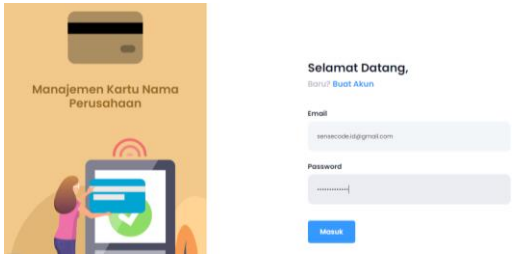
Selain mengubah data profil, pengguna juga dapat mengubah password dengan menekan tombol menu password di halaman pengaturan profil. Saat akan mengubah password, pengguna haru menginputkan data password lama, password baru, dan konfirmasi password baru. Setelah memasukkan data password, pengguna dapat menyimpan passwordnya dengan menekan tombol simpan di bagian bawah form ubah password.



Gambar 12 Tampilan Mengubah Password

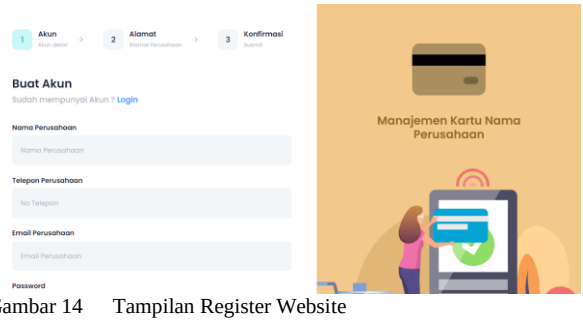
- Interface Login Register Admin

Halaman login register admin akan digunakan oleh pengguna perusahaan untuk masuk ke sistem ini melalui website. Untuk dapat masuk ke sistem, pengguna harus memasukkan email dan juga password dari akun perusahaan yang sudah didaftarkan sebelumnya. Tampilan login dari website dapat dilihat pada Gambar 13.



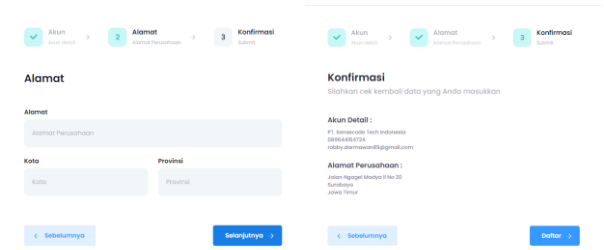
Gambar 13 Tampilan Login Website

Seperti pada Gambar 13, untuk pengguna baru yang ingin mendaftarkan perusahaannya, pengguna dapat menekan tombol buat akun yang berada di bagian tengah atas tampilan. Sedangkan untuk masuk, pengguna hanya perlu memasukkan email dan password lalu menekan tombol masuk yang ada pada halaman login.



Gambar 14 Tampilan Register Website

Ketika pengguna menekan tombol buat akun, maka pengguna akan dialihkan ke halaman register seperti pada Gambar 14. Form pengisian data perusahaan akan berupa wizard. Di wizard yang pertama, pengguna diminta untuk memasukkan data akun perusahaan mulai dari nama perusahaan, telepon perusahaan, email perusahaan, dan password untuk masuk ke sistem ini.

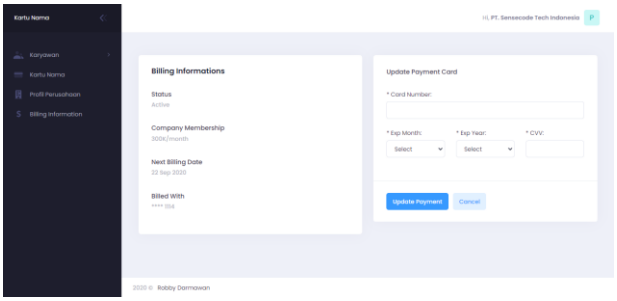


Gambar 15 Tampilan Register Website (2)

Pada wizard kedua yang dapat dilihat pada Gambar 15, pengguna diminta untuk memasukkan data alamat dari perusahaan yang akan didaftarkan. Setelah itu, pengguna akan dialihkan ke wizard ketiga untuk mengkonfirmasi data yang dimasukkan telah benar. Setelah perusahaan berhasil mendaftarkan perusahaannya, perusahaan akan menerima email verifikasi sesuai dengan email yang telah dimasukkan sebelumnya. Setelah pengguna melakukan email verifikasi, pengguna akan langsung diarahkan ke halaman website perusahaan.

- Interface Pengaturan Langganan

Pengguna perusahaan, dapat melakukan pembayaran biaya langganan menggunakan kartu kredit atau kartu debit yang memiliki visa atau mastercard. Di halaman pengaturan langganan ini, pengguna dapat memasukkan informasi kartu mereka mulai dari nomor kartu, tanggal kadaluarsa kartu, dan juga cvv kartu. Selain itu, di bagian sebelah kiri seperti pada Gambar 16, pengguna juga dapat melihat informasi biaya dan tanggal penagihan selanjutnya.



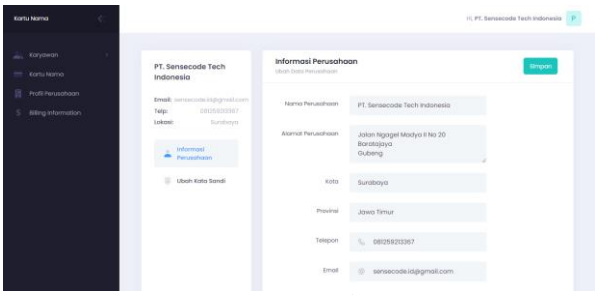
Gambar 16 Tampilan Pengaturan Langganan

- Interface Pengaturan Profil Perusahaan

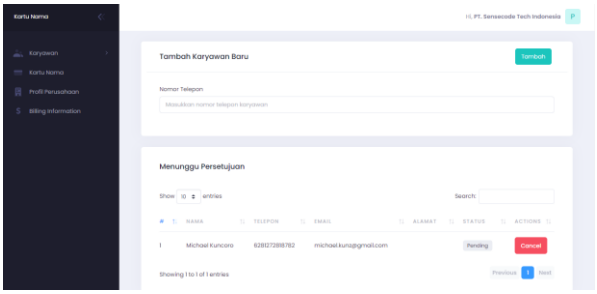
Perusahaan dapat mengatur profilnya di menu profil perusahaan. Ada dua submenu yang dapat dipilih oleh pengguna yakni, submenu informasi perusahaan dan submenu ubah kata sandi. Seperti pada Gambar 17, pengguna dapat mengubah informasi perusahaan pada input teks dan menekan tombol simpan pada bagian kanan atas. Sedangkan untuk mengubah kata sandi, pengguna harus memasukkan kata sandi lama terlebih dahulu untuk dapat mengubahnya menjadi kata sandi yang baru.

- Interface Karyawan Perusahaan

Tampilan karyawan perusahaan, merupakan tampilan yang digunakan akun perusahaan untuk mengatur data karyawan yang bergabung ke akun perusahaannya tersebut. Seperti pada Gambar 18, perusahaan dapat menambahkan karyawan dengan memasukkan nomor telepon karyawan dan menekan tombol tambah pada bagian kanan atas. Setelah ditambahkan, status karyawan akan pending hingga karyawan tersebut menerima undangan bergabung dari perusahaan tersebut.



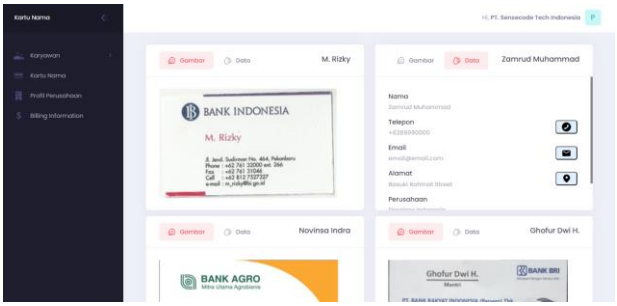
Gambar 17 Tampilan Pengaturan Profil



Gambar 18 Tampilan Karyawan Perusahaan

- Interface Kartu Nama Perusahaan

Setiap karyawan yang sudah bergabung dengan perusahaan dapat berkontribusi untuk menambahkan data kartu nama perusahaan yang baru. Daftar kartu nama tersebut akan muncul pada halaman kartu nama seperti pada Gambar 19. Pada tampilan tersebut dapat dilihat bahwa ada dua menu untuk masing-masing kartu nama. Menu yang pertama adalah menu gambar yang akan menampilkan gambar dari kartu nama yang disimpan oleh karyawan. Sedangkan menu kedua adalah menu data yang akan menampilkan seluruh informasi dari kartu nama tersebut. Untuk memudahkan pengguna dalam memanfaatkan informasi tersebut, pengguna dapat melakukan beberapa aksi pada menu data. Beberapa aksi tersebut adalah melakukan panggilan telepon, mengirim email, dan yang terakhir adalah membuka peta sesuai dengan alamat yang tertera pada kartu nama.



Gambar 19 Tampilan Kartu Nama Perusahaan

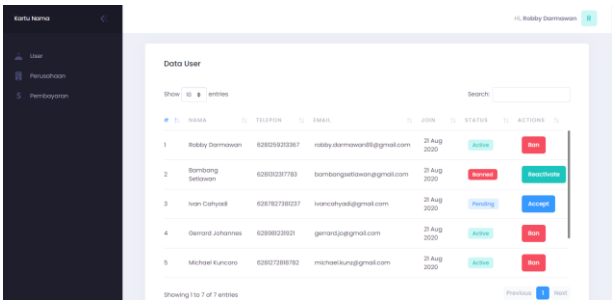
- Interface Manajemen Pengguna

Tampilan manajemen pengguna ini akan digunakan oleh super admin untuk mengatur akun pengguna. Super admin dapat melihat daftar pengguna individu maupun perusahaan pada menu yang ada. Dapat dilihat pada Gambar 20, super admin dapat melakukan beberapa aksi seperti memblokir pengguna, menerima aktivasi secara

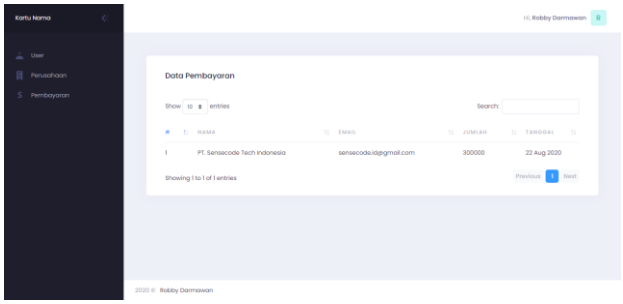
manual, dan mengaktifkan ulang pengguna yang telah diblokir sebelumnya.

- Interface Riwayat Pembayaran

  Seperti yang dijelaskan sebelumnya, pengguna perusahaan akan dikenakan biaya langganan setiap bulannya. Oleh karena itu, super admin dapat melihat riwayat dari pembayaran biaya langganan pada menu riwayat pembayaran. Tampilan riwayat pembayaran dapat dilihat pada Gambar 21.



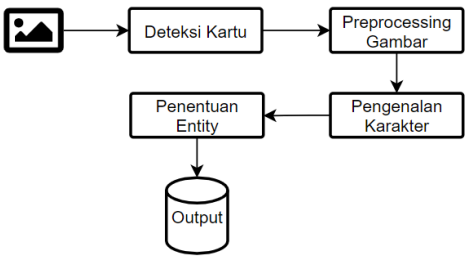
Gambar 20 Tampilan Manajemen Pengguna



Gambar 21 Tampilan Riwayat Pembayaran

IV. EKSTRAKSI INFORMASI KARTU NAMA

Pada bagian ini akan dijelaskan tentang proses ekstraksi dari pengenalan informasi pada kartu nama yang dibuat untuk penelitian ini. Proses ekstraksi informasi pada kartu nama ini akan dibahas akan dibagi menjadi 5 bagian yaitu implementasi untuk pengenalan kartu, preprocessing gambar, pengenalan karakter, penentuan entity kartu nama dan implementasinya pada webservice flask. Gambar 22 adalah arsitektur dari ektraksi informasi kartu nama pada Penelitian ini.

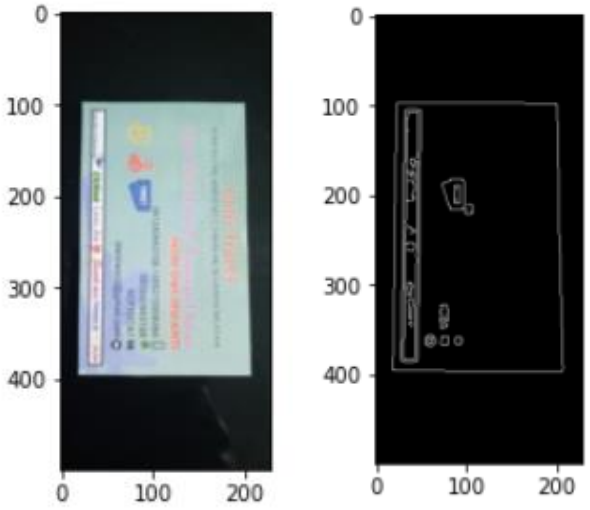


Gambar 22 Arsitektur Ekstraksi Kartu Nama

Dalam proses ekstraksi ini, sistem akan memiliki input berupa gambar dan mengeluarkan output berupa array of object. Array of object ini adalah objek dengan pasangan data beserta tipe entity hasil dari ekstraksi kartu nama. Gambar yang menjadi input pada sistem ini akan melewati 4 proses utama yakni, pendeteksian kartu, preprocessing gambar, pengenalan karakter, dan penentuan entity. Masing-masing proses akan dijelaskan satu per satu secara detail dengan potongan program dan gambar hasil dari masing-masing proses.

A. Pengenalan Kartu

Saat pertama kali webservice menerima gambar dari program, webservice akan memotong terlebih dahulu bagian background dari gambar yang bukan merupakan bagian dari kartu nama. Hal ini dilakukan dengan mendeteksi bentuk kotak pada gambar. Pendeteksian tepi dari objek gambar dilakukan dengan menggunakan metode canny. Input dan output dari hasil pendeteksian tepi ini dapat dilihat pada Gambar 23.



Gambar 23 Input Ouput Deteksi Tepi dengan Canny

B. Preprocessing Gambar

Setelah mendapatkan potongan kartu nama, program akan melakukan preprocessing sebelum ocr dilakukan. Hal ini tentunya bertujuan untuk memperoleh hasil ocr yang lebih baik. Preprocessing yang dilakukan dimulai dari perbaikan orientasi gambar dan membuang noise pada gambar.

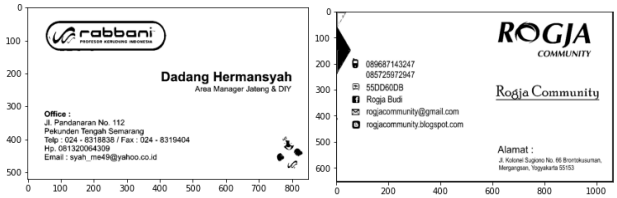
Pertama, program akan mendapatkan nilai orientasi gambar dengan memanfaatkan fungsi pada pytesseract yakni image_to_osd. Fungsi ini akan mengembalikan string data dari orientasi gambar. Oleh karena itu, selanjutnya digunakan regex untuk mendapatkan nilai orientasi dari string yang telah didapatkan sebelumnya. Selanjutnya, program akan menggunakan fungsi rotate dari ndimage untuk memutar gambar sehingga gambar akan memiliki orientasi yang tepat. Setelah memiliki orientasi yang tepat, program akan melakukan beberapa perbaikan pada gambar, semisalnya menghilangkan beberapa noise dengan melakukan thresholding. Hasil dari preprocessing gambar ini dapat dilihat pada Gambar 24.

Dalam preprocessing gambar ini, dilakukan beberapa proses untuk mengolah gambar sehingga gambar yang akan dijalankan pada proses pengenalan karakter dengan tesseract OCR menjadi lebih maksimal. Beberapa faktor seperti terang gelapnya warna background dan juga background yang bertekstur akan dihilangkan pada tahap preprocessing gambar ini.

Gambar 24 Hasil Preprocessing Kartu Nama

Gambar 25 Contoh Variasi Kartu Nama

Gambar 25 adalah contoh variasi dari gambar kartu nama. Dari gambar tersebut, dapat dilihat bahwa desain dari kartu nama sangat beragam baik dari pewarnaan dan juga tipe font. Gambar 25 bagian kiri merupakan contoh kartu nama yang memiliki warna background gelap dan bagian kanan merupakan contoh kartu nama yang memiliki warna background terang namun memiliki tekstur background yang bisa saja mengganggu proses pengenalan karakter yang dilakukan oleh tesseract OCR.

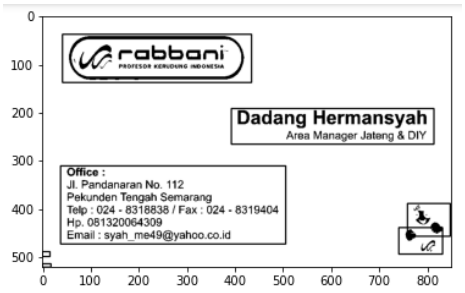


Gambar 26 Hasil Preprocessing Kartu Nama

Gambar 26 adalah hasil dari preprocessing Gambar 25. Tahap ini akan melakukan thresholding terhadap background yang ada sehingga warna background yang mengganggu proses pengenalan karakter dapat dihilangkan. Selain itu, untuk gambar yang memiliki warna background yang gelap, akan dibalik nilainya sehingga warna background akan menjadi warna yang terang (putih).

C. *Pengenalan Karakter*

Setelah gambar kartu nama selesai melalui proses preprocessing, maka gambar akan dilanjutkan ke proses ocr memanfaatkan tesseract [9]. Proses ocr pada Penelitian ini akan memanfaatkan library pytesseract dan mengikuti pendekatan pada [10] dan [11]. Sebelum melalui proses ocr, gambar yang sudah mengalami preprocessing akan dibagi-bagi sesuai kontur yang ada seperti pada Gambar 27. Hal ini bertujuan untuk memisahkan tulisan jika ada gambar yang memiliki desain informasi beberapa kolom.



Gambar 27 Pengelompokkan Kontur pada Kartu Nama

Gambar 27 merupakan contoh hasil dari pengelompokkan kontur pada gambar kartu nama. Hal ini dilakukan untuk dapat memisahkan kata ketika terdapat informasi atau tulisan yang terdiri dari beberapa kolom pada satu baris. Dari masing-masing kontur tersebut, akan dipotong dan dilakukan pengenalan karakter satu persatu.

D. *Penentuan Entity Kartu Nama*

Array variabel hasil proses ocr sebelumnya akan diklasifikasikan satu per satu manfaatkan teknik information extraction pada [12]. Pengklasifikasian entity akan memiliki dua macam klasifikasi. Klasifikasi yang pertama berdasarkan aturan yang dibuat dan yang kedua menggunakan model yang telah dibuat. Untuk klasifikasi yang pertama, hanya akan menghasilkan klasifikasi entity dengan tipe nomor telepon, email, dan website. Klasifikasi ini akan menggunakan regular expression untuk dapat mengklasifikasikan entity yang diinginkan.

E. *Implementasi Webservice Flask*

Pada bagian ini akan dijelaskan mengenai pengimplementasian program-program yang sudah dijelaskan sebelumnya ke dalam webservice. Untuk mengimplementasikan seluruh program yang telah dijelaskan sebelumnya, sistem membutuhkan sebuah framework yang dapat menjalankan bahasa pemrograman python dan juga menjadi sebuah service yang dapat dipanggil oleh aplikasi mobile.

Sistem manajemen kartu nama pada Penelitian ini akan memanfaatkan framework Flask sebagai webservice ekstraksi informasi pada gambar kartu nama. Karena framework ini menggunakan bahasa python maka sebelum menggunakan framework ini tentunya terlebih dahulu harus menginstal python dan pip pada perangkat yang akan digunakan.

Setelah program Flask berhasil dibuat, tentunya perlu ada beberapa baris program yang dibuat supaya webservice Flask dapat dipanggil dari aplikasi kita. Setiap potongan program yang telah dibuat sebelumnya, akan dijadikan satu pada file app.py di dalam proyek Flask yang telah dibuat. Masing-masing program akan dimasukkan ke dalam fungsi sehingga mudah dipanggil oleh main program yang akan menjadi webservice pada sistem Penelitian ini.

Data yang dikembalikan akan berupa array yang masing-masing merupakan objek data dan entity dari hasil ekstraksi.

Selain itu, webservice juga akan mengembalikan data gambar yang telah diproses dan diubah dalam bentuk base64.


```
▸1: {data: "Novinsa Indra", tag: "name"}
▸2: {data: "Kepala Divisi Bisnis Komersial", tag: "job_title"}
▸3: {data: "PT Bank Agroniaga Tbk.", tag: "company"}
▸4: {data: "Piaza GRI Lantai 16", tag: "name"}
▸5: {data: "Jl HR Rasuna Said Blok X2 No 1", tag: "address"}
▸6: {data: "Jakarta 12950 Indonesia", tag: "company"}
▸7: {data: "0215262570", tag: "phone"}
▸8: {data: "0215262653", tag: "phone"}
▸9: {data: "0811305467", tag: "phone"}
▸10: {data: "novinsa@bankagro.co.id", tag: "email"}
img: "iVBORw0KGgoAAAANSUhEUgAAArwAAAGrCAIAAACt49OqAAEAAE Show more (238 kB) Copy "
```


Gambar 28 Response Data Ekstraksi Kartu Nama

V. UJI COBA

Pada bagian ini akan dijelaskan mengenai uji coba yang dilakukan pada Penelitian Sistem Manajemen Kartu Nama dengan Ekstraksi Informasi Kartu Nama Otomatis. Uji coba merupakan tahap dalam pembuatan sebuah sistem untuk dapat mengetahui fungsionalitas dan kemampuan yang dimiliki oleh sistem yang dibuat. Uji coba akan mencakup pengujian fungsionalitas dengan metode blackbox testing dan usability dari penggunaan sistem yang telah dibuat. Selain itu juga akan dilakukan uji coba untuk mengetahui akurasi dari ekstraksi informasi kartu nama.

A. *Uji Coba Fungsionalitas*

Pada bagian ini akan dijelaskan mengenai uji coba yang dilakukan oleh developer secara langsung. Uji coba akan dilakukan dengan menggunakan blackbox testing. Uji coba ini dilakukan untuk mengetahui apakah fungsionalitas aplikasi berjalan dengan baik dan sesuai dengan yang dirancangkan pada bab sebelumnya. Uji coba ini akan dilakukan untuk masing-masing fitur yang ada pada aplikasi. Dari hasil uji coba ini, sistem yang dibuat telah memberikan output sesuai target yang diharapkan untuk setiap aksi yang dilakukan oleh pengguna.

B. *Uji Coba Usability*

Pada bagian ini akan dipaparkan hasil dari uji coba sistem yang dilakukan oleh pengguna secara langsung. Uji coba dilakukan kepada 25 orang yang akan mencoba menggunakan sistem sebagai user individu dengan aplikasi mobile dan 7 orang akan mencoba menggunakan sistem sebagai user perusahaan dengan aplikasi website. Dalam uji coba ini, aplikasi juga mengalami uji coba compatibility dimana aplikasi dijalankan setidaknya pada 7 tipe handphone yang berbeda seperti Vivo S1, Redmi Note 7, Huawei Honor Play, Oppo F7, dsb. Setelah mencoba semua fitur yang ada pada aplikasi, tester diberikan sebuah kuesioner untuk mengetahui response dari sisi pengguna menggunakan Google Form.

Tabel II merupakan hasil kuesioner dari pertanyaan yang telah diajukan kepada tester setelah mencoba seluruh fitur pada aplikasi dengan skala penilaian dari 1 sampai 4. Nilai 1 merupakan nilai yang paling rendah dan 4 merupakan nilai yang paling tinggi. Pertanyaan pada nomor 1 hingga nomor 7 merupakan pertanyaan yang diajukan kepada tester yang telah mencoba menggunakan aplikasi mobile sebagai pengguna individu pada sistem ini. Untuk pertanyaan nomor 8 hingga nomor 14 merupakan pertanyaan yang diajukan kepada tester yang menggunakan website sebagai admin perusahaan pada sistem ini.

C. *Uji Coba Akurasi*

Pada bagian ini akan dijelaskan hasil dari uji coba fitur ekstraksi informasi kartu nama. Uji coba ini akan menghitung akurasi dari proses pengenalan karakter dan juga klasifikasi entity dari kartu nama tersebut. Uji coba ini akan menghasilkan persentase akurasi dari pengujian ekstraksi kartu nama. Berikut ini adalah beberapa contoh dari proses uji coba yang dilakukan.

TABEL II
HASIL KUESIONER

Pertanyaan	1 Poin	2 Poin	3 Poin	4 Poin
Bagaimana tampilan aplikasi dari Sistem Manajemen Kartu Nama?	0%	4%	28%	68%
Bagaimana performa aplikasi yang disajikan?	0%	0%	60%	40%
Kemudahan dalam menggunakan aplikasi	0%	4%	48%	48%
Kemudahan sistem navigasi pada aplikasi	0%	0%	44%	56%
Kemudahan dalam menambahkan kartu nama	0%	8%	48%	44%
Ketepatan ekstraksi kartu nama otomatis	0%	20%	48%	32%
Apakah anda merasa terbantu untuk mengelola kartu nama dengan aplikasi ini?	0%	4%	48%	48%
Bagaimana tampilan website dari Sistem Manajemen Kartu Nama?	0%	0%	14.3%	85.7%
Bagaimana performa website yang disajikan?	0%	0%	57.1%	42.9%
Kemudahan dalam menggunakan website	0%	0%	28.6%	71.4%
Kemudahan sistem navigasi pada website	0%	0%	57.1%	42.9%
Kemudahan dalam mengelola karyawan perusahaan	0%	0%	42.9%	57.1%
Kemudahan dalam mengelola langganan	0%	0%	28.6%	71.4%
Apakah anda merasa terbantu untuk mengelola kartu nama karyawan dengan sistem ini?	0%	10%	50%	40%

Gambar 29 Contoh Input 1 Kartu Nama

Tabel III merupakan hasil uji coba dari pengenalan karakter dan klasifikasi entity yang dihasilkan oleh sistem dengan menerima input seperti Gambar 29. Contoh input dan output tersebut merupakan salah satu uji coba yang memiliki akurasi yang tinggi baik dalam pengenalan karakter dan juga pengklasifikasian entity. Hampir seluruh data pada kartu nama Gambar 29 dapat diekstrak dengan baik.

Uji coba untuk proses pengenalan karakter ini akan menghitung akurasi dari setiap karakter yang dihasilkan. Pada pengujian ini, tingkat kesalahan pengenalan karakter akan dihitung menggunakan metode levensthein distance.

Levensthein distance adalah algoritma yang mengukur kesamaan antara 2 string. Dalam pengujian pengenalan karakter ini, pengujian dilakukan menggunakan 30 macam kartu nama yang dapat dilihat pada lampiran A. Dari pengujian ini terdapat 5028 karakter dengan tingkat kesalahan 749. Dengan perhitungan (Total Karakter – Kesalahan) / Total Karakter * 100%, maka dapat disimpulkan bahwa pengenalan karakter dari sistem ini memiliki akurasi sebesar 85.1%. Sedangkan untuk penghitungan akurasi pada klasifikasi entity menggunakan rata-rata dari F1 score yang diperoleh dari nilai precision dan recall untuk masing-masing entity. Dari perhitungan tersebut, akurasi klasifikasi pada sistem kartu nama ini adalah 86%.

TABEL III
OUTPUT 1 KARTU NAMA

No	Output Pengenalan Karakter	Output Klasifikasi Entity
1	BANK AGRO	Company
2	Novinsa Indra	Name
3	Kepala Divisi Bisnis Komersial	Job Title
4	PT Bank Agroniaga Tbk.	Company
5	Plaza GRI Lantai 16	Address
6	Jl. HR. Rasuna Said Blox X-2, No. 1	Address
7	Jakarta 12950, Indonesia	Company
8	0215262570	Phone
9	0215262653	Phone
10	0811305467	Phone
11	novinsa@bankagro.co.id	Email

VI. KESIMPULAN

Pada bagian ini akan dijelaskan mengenai kesimpulan yang diperoleh setelah pembuatan Sistem Manajemen Kartu Nama dengan Ekstraksi Informasi Kartu Nama Otomatis. Berikut adalah kesimpulan yang telah diperoleh dalam pembuatan sistem ini.

- Ekstraksi Informasi Otomatis dapat mempermudah pengguna dalan menyimpan kartu nama.
- Perusahaan dapat memanajemen kartu nama dengan mudah dan tidak perlu khawatir kehilangan data kartu nama dari karyawan perusahaan.
- Microframework Flask merupakan solusi mudah dalam membuat webservice yang membutuhkan bahasa pemrograman Python beserta library-library Python.
- Klasifikasi data yang menggabungkan klasifikasi berbasis aturan dan klasifikasi menggunakan model Naive Bayes dapat menghasilkan ekstrak informasi kartu nama dengan cukup baik.
- Hasil uji coba pengenalan karakter pada sistem Penelitian ini memiliki akurasi sebesar 85.1%.
- Hasil uji coba pengklasifikasian entity pada sistem Penelitian ini memiliki akurasi sebesar 86%.

DAFTAR PUSTAKA

[1] Santoso, Sugeng, Josch Kauf, and Nabila Cynthia Aristo. "The Information System of Name Card Sales Based on Digital Marketing to Improve Creativepreneur on College E-Commerce Website." Aptisi Transactions On Technopreneurship (ATT) 1.1 (2019): 64-72.

[2] Vuong, Bao Quoc, and Hung Ngoc Do. "Design and implementation of multilanguage name card reader on Android platform." 2014 International Conference on Advanced Technologies for Communications (ATC 2014). IEEE, 2014.

[3] Yener, M., Dundar, O. 2016. *Expert Android Studio.* Wrox.

[4] He, Ren Yu. "Design and implementation of web based on Laravel framework." 2014 International Conference on Computer Science and Electronic Technology (ICCSET 2014). Atlantis Press, 2015.

[5] Redmond, Paul. "Lumen Programming Guide." New York: Penerbit Apress Media (2016).

[6] Dwyer, Gareth. 2016. *Flask By Example.* Packt Publishing.

[7] Smith, Ray. "An overview of the Tesseract OCR engine." Ninth international conference on document analysis and recognition (ICDAR 2007). Vol. 2. IEEE, 2007.

[8] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.

[9] Smith, Ray W. "History of the Tesseract OCR engine: what worked and what didn't." Document Recognition and Retrieval XX. Vol. 8658. International Society for Optics and Photonics, 2013.

[10] Thakare, Sahil, et al. "Document Segmentation and Language Translation Using Tesseract-OCR." 2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS). IEEE, 2018.'

[11] Warsito, Amelia Alexandra Putri, and Herry Pieter. "Aplikasi Manajemen Hutang Piutang dengan OCR Berbasis Swift." Journal of Intelligent System and Computation 2.2 (2020): 47-55.

[12] Ferdinandus, F. X., et al. "Optical Character Recognition Dan Information Extraction Pada Dokumen Kamus Bilingual." Inov. dalam Desain dan Teknol (2011): 302-308.

JOURNAL OF INTELLIGENT SYSTEMS AND COMPUTATION

Klasifikasi Keluhan Masyarakat Terhadap Layanan Publik pada Harian Radar Tarakan

Indra Tri Saputra, Program Studi Sistem Informasi, STMIK PPKIA Tarakanita Rahmawati Tarakan.

***Abstract*− *Website* koran harian Radar Tarakan memiliki kolom dengan judul "Warga Menulis" di mana menu ini merupakan sarana bagi pembaca untuk menyampaikan keluhan ataupun aspirasi mereka. Yang menjadi permasalahan, pesan pembaca atau opini yang ditampilkan bersifat *to the point,* hanya isi opini sesuai yang dikirim pembaca tanpa informasi tambahan kepada siapa opini tersebut ditujukan. Tujuan dari penelitian ini adalah melakukan klasifikasi data opini pada *website* koran harian Radar Tarakan khususnya opini yang berkaitan dengan fasilitas dan pelayanan publik. Klasifikasi merupakan suatu proses pengelompokkan data sesuai dengan kelas atau kategori yang telah ditentukan sebelumnya. Hipotesis yang dapat diambil adalah hasil klasifikasi diharapkan memiliki akurasi hingga 70%. Tahap awal dari proses klasifikasi yaitu *preprocessing* di mana pada tahap ini hal-hal yang dilakukan antara lain *case folding, tokenizing, convert word, stopword removal (filtering)* dan *stemming*. Algoritma yang digunakan dalam penelitian ini adalah Frequency Ratio Accumulation Method (FRAM). Pembuatan aplikasi menggunakan bahasa pemrograman PHP dan *database* MySQL. Hasil uji coba dari penelitian ini menunjukkan rata-rata akurasi yang diperoleh pada proses klasifikasi opini menggunakan algoritma FRAM adalah 60%. Besar kecilnya prosentase akurasi tergantung dari jumlah data latih yang digunakan. Semakin banyak jumlahnya dapat meningkatkan nilai akurasi akan tetapi hal ini akan berpengaruh terhadap efisiensi kinerja sistem.**

***Kata Kunci*− FRAM, Klasifikasi, Opini, Machine Learning**

I. PENDAHULUAN

Pemanfaatan teknologi informasi khususnya internet perlu dilakukan secara maksimal guna memenuhi kebutuhan masyarakat akan informasi yang up to date. Kebutuhan masyarakat akan informasi secara cepat dan mudah mendorong media-media penyedia informasi menyajikan informasi secara online. Salah satu yang memanfaatkan kelebihan internet tersebut adalah surat kabar harian atau koran.

Koran harian Radar Tarakan memiliki website dengan alamat www.kaltara.procal.co. Seperti website surat kabar pada umumnya, secara keseluruhan menu yang ada pada website www.kaltara.procal.co berisi informasi berita baik lokal, nasional, internasional maupun informasi penting lainnya.

Indra Tri Saputra, Program Studi Sistem Informasi, STMIK PPKIA Tarakanita Rahmawati Tarakan, Indonesia (e-mail: indratriputra@gmail.com)

Selain informasi berita terdapat satu kolom dengan judul "Warga Menulis" di mana menu ini merupakan sarana bagi pembaca untuk menyampaikan keluhan ataupun aspirasi mereka khususnya yang berhubungan dengan fasilitas dan pelayanan publik.

Pesan pembaca atau opini yang ditampilkan bersifat to the point, hanya isi opini sesuai yang dikirim pembaca tanpa informasi tambahan kepada siapa opini tersebut ditujukan. Sementara tidak semua opini secara terbuka menyertakan hal tersebut. Atau bisa saja pembaca keliru mengalamatkan opininya. Padahal opini berisi hal-hal yang ingin disampaikan pembaca mengenai permasalahan yang dihadapi dengan harapan akan ada penyelesaian atau tanggapan untuk permasalahan tersebut. Belum adanya penekanan kepada siapa opini tersebut ditujukan bisa berakibat opini tidak akan tersampaikan dengan benar. Untuk itu perlu adanya penekanan yang jelas agar keluhan pembaca tidak salah sasaran.

Sebuah disiplin ilmu text mining dapat digunakan untuk mengatasi permasalahan tersebut. Text mining [1] adalah suatu proses untuk mengambil informasi dari teks yang ada. Text mining mencari pola-pola yang ada di teks dalam bahasa natural yang tidak terstruktur. Algoritma yang digunakan dalam penelitian ini adalah Frequency Ratio Accumulation Method (FRAM). Algoritma FRAM merupakan salah satu metode pembelajaran supervised document classification. Metode klasifikasi F*RAM* adalah metode klasifikasi yang menggunakan jumlah rasio frekuensi pada tiap kategori dari fitur term individual.

Langkah yang akan diambil pada penelitian di awali dengan memasukkan opini dari alamat website koran harian Radar Tarakan www.kaltara.procal.co ke dalam database, kemudian dilakukan tahapan preprocessing. Langkah selanjutnya melakukan perhitungan menggunakan metode FRAM untuk memperoleh klasifikasi opini yang tepat sebelum ditampilkan pada alamat website www.kaltara.procal.co. Hasil akhir penelitian ini nantinya secara otomatis akan menampilkan opini berikut kelas atau kategorinya berdasarkan isi dari opini sesuai klasifikasi yang sudah tersedia.

II. TINJAUAN PUSTAKA

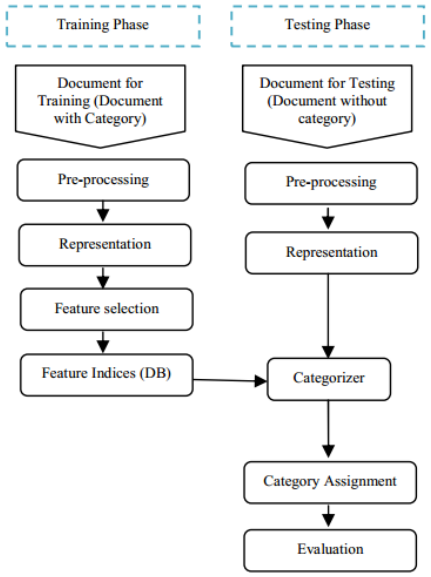
Kategorisasi teks [2] berkaitan dengan area informasi dan dokumentasi pengetahuan. Karena informasi dan pengetahuan disimpan dan dibagi ke dalam beberapa kategori atau teks, kategorisasi teks membantu pengguna dalam hal informasi dengan mengarahkan ke informasi yang diinginkan. Sebagian besar teknik kategorisasi teks seperti Machine Learning [3], [4],

Naïve Bayes (NB), K-Nearest Neighbour (KNN) [5], Support Vector Machine (SVM) dan Decision Tree memiliki model matematika yang kompleks [6]. Semuanya cenderung berkaitan dengan seleksi fitur dan categorization successively menjadi permasalahan tersendiri dalam kategorisasi teks otomatis, yang menyebabkan biaya dan masalah komputasi yang kompleks [7], [8].

Frequency Ratio Accumulation Method (FRAM) merupakan teknik klasifikasi sederhana yang menambahkan rasio dari frekuensi term pada masing-masing kategori, dan bisa menggunakan indeks term tanpa batas [9]–[12].

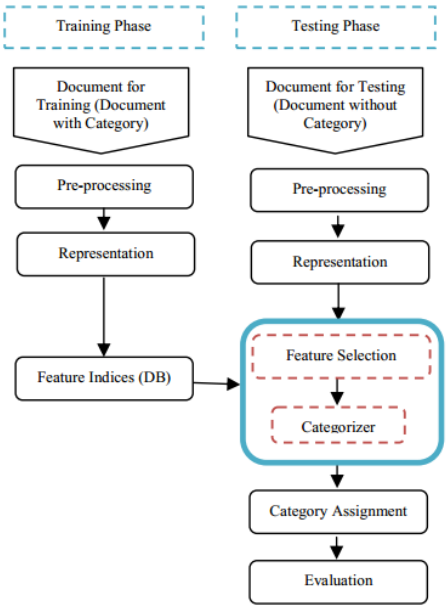
Metode ini ditandai sebagai klasifikasi dokumen tanpa mengekstraksi fitur term pada tahap seleksi fitur. Tugas kategorisasi dikombinasikan dengan tugas pengolahan fitur. Perbedaan tahap-tahap main building blocks antara kategorisasi teks otomatis standar dengan kategorisasi teks otomatis metode FRAM ditunjukkan pada Gambar 1 dan Gambar 2.

Tahapan proses baik itu pada kategorisasi teks otomatis standar maupun kategori teks otomatis FRAM terdiri dari 2 (dua) proses yaitu yang pertama adalah tahap pelatihan dimana dokumen yang diberi label di bawah kategori yang telah ditetapkan awalnya dilakukan inisialisasi pra-proses untuk menghilangkan term yang tidak berguna dan mengganggu. Selanjutnya, fitur term yang penting menjadi kata kunci diekstrak dalam tahap pelatihan mulai representasi hingga proses seleksi fitur dan menghasilkan database indeks, selanjutnya disebut Database (DB), yang kemudian digunakan untuk tahap yang kedua yaitu tahap uji. Pada tahap uji, pemeriksaan klasifikasi akan dievaluasi dengan mengelompokkan satu set dokumen pra-kategori satu demi satu sebagai dokumen tanpa kategori, dan kemudian mengukur kinerja kategorisasi menggunakan beberapa teknik standar evaluasi kinerja.



Gambar. 1. Kategorisasi Teks Otomatis Standar

Perbedaan antara kategorisasi teks otomatis FRAM pada Gambar 2 dengan kategori teks otomatis standar pada Gambar 1 terjadi pada tahap pertama yaitu tahap pelatihan dimana pada kategorisasi teks otomatis FRAM proses seleksi fitur akan dikecualikan dan proses seleksi fitur akan ditetapkan kategorinya pada tahap klasifikasi.



Gambar. 2. Kategorisasi Teks Otomatis FRAM

Hal ini yang menjadi kelebihan dari kategorisasi teks otomatis FRAM dimana dengan tidak adanya proses seleksi fitur pada tahap pelatihan akan mengurangi waktu komputasi yang berdampak pada kinerja sistem menjadi lebih efisien.

Metode ini pada awalnya menghitung total rasio frekuensi (FR) fitur term individual dalam tiap kategori sebagai berikut:

$$FR(t_n, c_k) = \frac{R(t_n,c_k)}{\sum_{c_k \in C} R(t_n,c_k)} \quad (1)$$

Di mana, rasio R tiap fitur term untuk masing-masing kategori dihitung dengan:

$$R(t_n, c_k) = \frac{f_{c_k}(t_n)}{\sum_{t_n \in T} f_{c_k}(t_n)} \quad (2)$$

$f_{c_k}(t_n)$:
$Total\ Frequency\ fitur\ t_n\ di\ dalam\ kategori\ c_k$

Dengan demikian, pada tahap pelatihan, rasio frekuensi dari semua fitur term dihitung dan dikumpulkan dalam tiap kategori. Selanjutnya, menghitung nilai evaluasi kategori atau skor kategori, di mana indikasi kemungkinan dokumen kandidat pada tahap uji termasuk dalam suatu kategori sebagai berikut:

$$E_{d_i}(c_k) = \sum_{t_n \in d_i} FR(t_n, c_k) \quad (3)$$

Akhirnya, dokumen kandidat d_i diklasifikasikan ke dalam kategori $c\hat{}_k$ di mana skor kategorinya maksimum, seperti berikut:

$$c_{\hat{k}} = \arg\max_{c_k \in C} E_{d_i}(c_k) \quad (4)$$

Metode yang diusulkan yaitu FRAM mempertahankan rasio frekuensi dalam tahap pelatihan dengan total jumlah fitur term yang di lambangkan dengan N dan total jumlah kategori yang di lambangkan dengan K. Selain itu, skor kategori untuk tiap kategori dihitung dengan menambahkan rasio frekuensi dari dokumen kandidat pada tahap uji mencakup fitur term dan golongan fitur term ke dalam kategori terkait yang mana skor evaluasinya maksimum.

III. METODE DAN INTI PENELITIAN

Algorithma dalam penelitian ini terdiri dari 2 (dua) tahap, yaitu tahap preprocessing data dan klasifikasi data opini.

A. *Praproses Data*

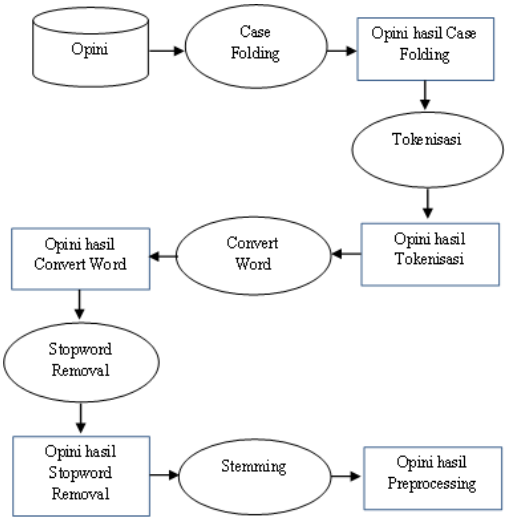
Pada tahap ini dilakukan serangkaian proses untuk mempersiapkan data. Seperti telah dijelaskan sebelumnya bahwa data opini yang digunakan masih berupa data teks tidak terstruktur (unstructured data), sehingga diperlukan langkah-langkah mempersiapkan data tersebut agar siap digunakan untuk proses selanjutnya.

1. Case Folding
   Tahap pertama adalah case folding yang bertujuan mengubah teks ke bentuk huruf kecil semua serta menghilangkan karakter seperti tanda baca dan angka.
2. Tokenisasi
   Tahap berikutnya tokenisasi yaitu memisahkan deretan kata pada kalimat, paragraf atau halaman menjadi token atau potongan kata tunggal (termmed word).
3. Convert Word
   Tahap selanjutnya convert word, yaitu merubah kata yang tidak baku menjadi lebih baku.
4. Stopword Removal (Filtering)
   Tahap berikutnya adalah stopword removal atau filtering, yaitu penghapusan term yang tidak berhubungan (irrelevant) dengan subyek utama dari database meskipun kata tersebut sering muncul dalam dokumen, misalnya: ada, adalah, agak, agar, akan dan sebagainya.
5. Stemming
   Tahap terakhir adalah menghilangkan awalan dan akhiran kata atau stemming. Dengan stemming setiap kata dikembalikan ke bentuk dasarnya.
   Adapun alur dari preprocessing data dapat dilihat pada Gambar 3.

B. *Klasifikasi Data Opini*

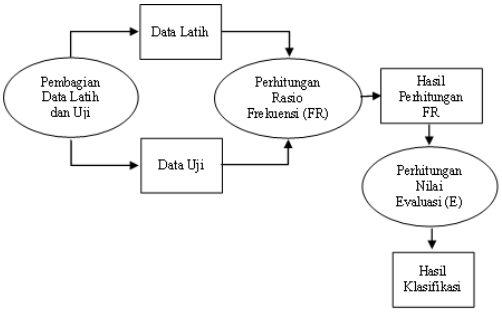
Klasifikasi menggunakan metode Frequency Ratio Accumulation Method (FRAM) melalui beberapa tahapan, yaitu:

1. Pembagian Data Latih dan Uji
   Setelah tahap praproses data (preprocessing), dataset dibagi menjadi 2 (dua) bagian, data latih (training) dan data uji (testing). Data latih digunakan sebagai data pembelajaran untuk menemukan pola yang tepat terhadap sekumpulan dataset, sehingga saat proses testing diperoleh hasil klasifikasi yang tepat. Pembagian data dilakukan dengan cara mengambil berapa prosen dataset untuk training dan berapa prosen untuk testing, misalnya 70% data training dan 30% data testing. Pembagian data dilakukan secara acak dari keseluruhan dataset yang digunakan.



Gambar. 3. Preprocessing Data

2. Perhitungan Rasio Frekuensi (FR)
   Kemudian data latih akan dibuat dalam bentuk matriks, di mana baris mewakili kata atau term dan kolom mewakili dokumen yang dikelompokkan berdasarkan kelas atau kategorinya masing-masing. Nilai matriks D(n,k) adalah jumlah kemunculan kata n pada kelas k, seperti ditunjukkan pada persamaan (2)
3. Perhitungan Nilai Evaluasi (E)
   Tahap akhir yaitu mencari nilai evaluasi (E) menggunakan persamaan (3)
   Adapun alur dari klasifikasi data opini menggunakan metode FRAM dapat dilihat pada Gambar 4.



Gambar. 4. Klasifikasi Opini Metode FRAM

C. *Metode Evaluasi*

Cross Validation merupakan salah satu teknik untuk menilai atau memvalidasi keakuratan sebuah model yang dibangun berdasarkan dataset tertentu. Dalam teknik ini dataset dibagi menjadi sejumlah K-buah partisi secara acak. Kemudian dilakukan sejumlah K-kali eksperimen, di mana masing-masing eksperimen menggunakan data partisi ke-K sebagai data testing dan memanfaatkan sisa partisi lainnya sebagai data training. Sebagai gambaran, jika kita melakukan 5-Fold Cross Validation maka desain data eksperimennya sebagai berikut.

Dataset: K1, K2, K3, K4, K5

TABLE I.
DATA EKSPERIMEN

Eksperimen Ke	***Data Testing***	***Data Latih***
1	K1	K2,K3,K4,K5
2	K2	K1,K3,K4,K5
3	K3	K1,K2,K4,K5
4	K4	K1,K2,K3,K5
5	K5	K1,K2,K3,K4

IV. HASIL EKSPERIMEN DAN PENELITIAN

Algoritma dalam penelitian ini diimplementasikan menggunakan bahasa pemrograman PHP dengan database MySQL. Prosesnya di awali dengan memasukkan opini dari alamat website koran harian Radar Tarakan www.kaltara.procal.co ke dalam database, kemudian dilakukan tahapan preprocessing. Langkah selanjutnya penentuan jenis kelas untuk data opini baru dengan menggunakan metode klasifikasi FRAM.

Perbandingan antara proses klasifikasi dengan cara manual ataupun menggunakan metode FRAM dengan jumlah dataset 500 opini belum mencapai hasil sebagaimana tertulis dalam hipotesis penelitian. Perbandingan antara proses klasifikasi manual dengan klasifikasi metode FRAM menggunakan teknik validasi *Cross Validation* rata-rata tingkat akurasi yang diperoleh adalah 51%.

Ketidaksesuaian antara kelas aktual dengan kelas prediksi pada proses klasifikasi lebih disebabkan pada tidak berimbangnya jumlah data yang ada pada setiap kelas atau kategori. Semakin banyak jumlahnya dapat meningkatkan nilai akurasi akan tetapi hal ini akan berpengaruh terhadap efisiensi kinerja sistem. Daftar beberapa hasil klasifikasi opini baru dapat dilihat pada tabel II

TABLE II.
KESESUAIAN JENIS KELAS UNTUK DATA OPINI BARU

No.	***Opini***	***Kelas***	***Kesesuaian Jenis Kelas***
1	KEPADA DKPP, kenapa pembersihan jalan dilakukan pada siang hari, menganggu pengguna jalan. Debunya di mana-mana	DKPP	Sesuai
2	DINAS perhubungan kabupaten Bulungan dan kota Tarakan, mohon diperhatikan perlengkapan keselamatan speedboat bagi penumpangnya. Bila perlu diperiksa terlebih dahulu perlengkapan-perlengkapan tersebut, demi keselamatan	DISHUB	Sesuai
3	TARAKAN makin kacau, daging asal Malaysia dilarang masuk tapi daging Indonesia harganya selangit. Minyak tanah susah ditiap RT tapi pengecer di sepanjang jalan selalu banyak dapat minyak tanah & dijual dengan harga 45 ribu tiap 5 liter. Apa pemerintah Tarakan tidak bisa mengendalikan semua ini untuk kesejahteraan masyarakat, khususnya Tarakan?!	DISPERINDAGKOP	Sesuai
4	Lagi Dinas PU Kabupaten Nunukan memperhatikan jalan dari Long Bawan sampai lagi ke Midang Kecamatan Krayan. Karena sangat rusak parah.	PEMDA	Berbeda
5	KEPADA perusahan yang sudah diberi surat edaran tentang UMK, segera naikan gaji karyawan. Jangan sampai memberi gaji karyawan tidak sesuai ketentuan	DISOSNAKER	Sesuai

V. KESIMPULAN

Dari percobaan yang dilakukan pada proses klasifikasi data opini masyarakat menggunakan metode FRAM, hasil yang diperoleh untuk akurasi terbesar atau tertinggi adalah 60%. Perbandingan antara proses klasifikasi manual dengan klasifikasi metode FRAM menggunakan teknik validasi *Cross Validation* rata-rata tingkat akurasi yang diperoleh adalah 51%.

Ketidaksesuaian antara kelas aktual dengan kelas prediksi pada proses klasifikasi lebih disebabkan pada tidak berimbangnya jumlah data yang ada pada setiap kelas atau kategori. Perbandingannya adalah jumlah data terbesar ada pada kelas atau kategori "PLN" dengan 124 data dan jumlah data terkecil ada pada kelas atau kategori "PENGADILAN" dengan 9 data. Sedangkan rata-rata jumlah data pada setiap kelas adalah 45 data.

DAFTAR PUSTAKA

[1] R. Feldman, J. Sanger, and others, *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.

[2] R. J. Mooney, "Machine Learning Text Categorization," *Austin Univ. Texas*, 2006.

[3] Fitria, Gunawan, and E. I. Setiawan, "Abstract Summarization Using Maximum Marginal Relevance and Vector Space Model."

[4] M. Suzuki, N. Yamagishi, T. Ishida, M. Goto, and S. Hirasawa, "On a new model for automatic text categorization based on vector space model," in *2010 IEEE International Conference on Systems, Man and Cybernetics*, 2010, pp. 3152–3159.

[5] A. Indriani, E. Novianto, and others, "Weight Adjusted K-Nearest Neighbor dan Minimum Spanning Tree untuk Information Retrieval System di Perpustakaan STMIK PPKIA Tarakanita Rahmawati Tarakan," in *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, 2013, vol. 1, no. 1.

[6] G. Pant and P. Srinivasan, "Learning to crawl: Comparing classification schemes," *ACM Trans. Inf. Syst.*, vol. 23, no. 4, pp. 430–462, 2005.

[7] J. Wang and A. An, "Classification Methods," 2005, pp. 144–149.

[8] A. Mahinovs, A. Tiwari, R. Roy, and D. Baxter, *Text classification method review*. 2007.

[9] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," 2006.

[10] M. Suzuki and S. Hirasawa, "Text categorization based on the ratio of word frequency in each categories," in *2007 IEEE International Conference on Systems, Man and Cybernetics*, 2007, pp. 3535–3540.

[11] M. Suzuki, T. Ishida, and M. Goto, "Refinement of index term set and improvement of classification accuracy on text categorization," in *2008 International Symposium on Information Theory and Its Applications*, 2008, pp. 1–6.

[12] B. T. Sharef, N. Omar, and Z. T. Sharef, "An automated arabic text categorization based on the frequency ratio accumulation.," *Int. Arab J. Inf. Technol.*, vol. 11, no. 2, pp. 213–221, 2014.

JOURNAL OF INTELLIGENT SYSTEMS AND COMPUTATION

Pengenalan Tulisan Pada Iklan Pinggir Jalan yang Melengkung Menggunakan Shape Context

Endang Setyati, *Departemen Informatika, Institut Sains dan Teknologi Terpadu Surabaya,*
Raymond Sugiarto, *Departemen Informatika, Institut Sains dan Teknologi Terpadu Surabaya*.

***Abstrak*— Membaca sebuah tulisan yang sama di bidang melengkung berbeda dengan di bidang datar, karena tulisan pada bidang melengkung bergantung pada permukaan bidang lengkungnya. Pada saat ini, banyak sekali tulisan pada iklan pinggir jalan yang ditempel pada bidang melengkung di sepanjang jalan. Tulisan yang digunakan berupa huruf dan angka, dengan berbagai macam background, bentuk dan warna yang diambil di pinggir jalan dengan menggunakan Farey Shape Context. Fitur Farey ini bergantung pada DSS (Digital Straight Line Segment) endpoint dan menggunakan pecahan Augmented Farey sequence. DSS endpoint ini dijadikan sebagai titik fitur atau feature point untuk menemukan shape context dari citra. DSS endpoint tersebut digunakan sebagai acuan *bounding box* yang akan digunakan sebagai object boundary yang dimana setiap sudutnya merupakan *reference point.* Untuk melakukan Binning Farey Rank, Augmented Farey Table (AFT) harus dibentuk terlebih dahulu berdasarkan Augmented Farey Sequence yang merupakan pengembangan dari Farey Sequence. Farey Sequence hanya meliputi pecahan dengan pembilang dan penyebut yang positif, sedangkan Augmented Farey Sequence meliputi pecahan dengan pembilang dan penyebut positif serta negatif. Pada penelitian ini digunakan 500 data iklan di pinggir jalan yang melengkung, dimana 70% digunakan sebagai data sample. Dari 70% data sample tersebut didapatkan ribuan karakter berupa huruf dan angka yang dijadikan data sample. Berdasarkan hasil uji coba penelitian yang dilakukan pada 500 Gambar dimana 30% sebagai data testing, maka hasil Farey Shape Context untuk mengenali tulisan berupa huruf dan angka pada iklan pinggir jalan yang melengkung mencapai akurasi benar 74.94% dan salah 25.06%.**

***Kata Kunci*—Augmented Farey Sequence, Farey Table, Object Recognition, Shape Context, Shape Matching.**

I. PENDAHULUAN

Berkembangnya teknologi yang ada pada dunia menyebabkan berkembangnya dunia IT. Salah satu pengembangan yang berhubungan dengan Gambar / Image yang berupa karakter huruf atau angka. Data gambar berupa karakter huruf atau angka ini harus mampu diproses dan diolah sehingga sistem dapat mengenali huruf / angka tersebut. Dari captcha, hingga iklan - iklan yang ada di pinggir jalan kota Surabaya.

Endang Setyati, Departemen Informatika, Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Jawa Timur, Indonesia (e-mail: endang@stts.edu)

Raymond Sugiarto, Departemen Informatika, Institut Sains dan Teknologi Terpadu Surabaya,, Surabaya, Jawa Timur, Indonesia (e-mail: raymondsugiarto@gmail.com)

Dengan menggunakan percobaan beberapa metode pengenalan gambar diharapkan tingkat keberhasilan pengenalan gambar dapat meningkat sehingga lebih akurat. Dengan ada proses pengenalan ini diharapkan dapat membantu membaca iklan yang ada di pinggir jalan yang ditempelkan pada bidang melengkung.

Iklan - iklan melengkung di pinggir jalan ini terdiri dari beberapa karakter yang berupa huruf dan angka yang akan lebih sulit terbaca dari satu sisi. Selain itu iklan tersebut juga terdiri dari banyak macam background dan bentuk tulisan serta berbagai warna tulisan.

Proses pengenalan tulisan pada iklan - iklan pinggir jalan ini menggunakan metode algoritma pada Image Processing yaitu Shape Context. Metode ini dikenal sebagai metode yang dapat digunakan untuk mengenali suatu bentuk yang telah diubah - ubah. Bentuk yang telah diubah - ubah ini akan diterapkan pada iklan - iklan yang melengkung di pinggir jalan. Iklan tersebut difoto kemudian akan dilakukan percobaan pada sistem agar sistem dapat membaca tulisan pada foto tersebut. Dengan bantuan preprocessing pada algoritma Image Processing seperti grayscale, remove background dan sebagainya, akan membantu sistem untuk memperjelas foto / gambar iklan yang melengkung agar Shape Context dapat bekerja lebih maksimal untuk mengenali tulisan pada iklan yang melengkung tersebut.

II. TINJAUAN PUSTAKA

Penelitian yang ditulis oleh Sanjoy Pratihar, Najima Begum (2016, Computer Science and Engineering Department) [1] ini mengusulkan pengenalan karakter tulisan tangan dengan menggunakan metode *Shape Context*. Bentuk adalah atribut visual yang penting dalam gambar. Pencocokan bentuk banyak digunakan dalam *Computer Vision* seperti pelacakan objek, pengambilan gambar, pendaftaran gambar, pengenalan karakter optik dan lain lain. Penelitian ini telah menyajikan teknik pembahaman Shape Context dengan mudah dan efisien dalam kaitannya dengan urutan Farey dan pecahan Farey.

Pada tahap pre-processing, segmen garis lurus (DDS) yang menentukan batas bentuk yang diekstrasi dengan menggunakan algoritma yang diberikan. Titik akhir dari segmen garis lurus digital yang menentukan batas objek diekstrasi. Prewiit tepi telah digunakan untuk menguji apakah piksel adalah piksel tepi atau tidak. Empat simpul dari persegi panjang area minimum yang meloncat dari titik - titik yang diekstrasi ini digunakan sebagai titik acuan untuk analisis Shape Context. Vektor dari

empat titik acuan mewakili pecahan urutan Farey yang diperbesar (F n), yang mencakup ruang 360o penuh (masing-masing titik referensi mencakup satu kuadran).

Lekuk dari titik akhir yang diekstrak yang berkenaan dengan titik referensi dipresentasikan dengan peringkat pecahan yang sesuai dengan urutan Farey. Dari distribusi peringkat Farey, diperoleh histogram yang memberikan objek Shape Context.

Setelah realisasi histogram normal yang sesuai dengan bentuk, digunakan metrik pengukuran jarak Chi-square untuk pelaporan kesamaan. Jarak Chi-square (CSD) dari dua histogram h1 dan h2 diberikan di sini di Persamaan berikut

$$dist(h_1,h_2)=\frac{1}{2}\sum_{i=0}^{n_b-1}\frac{(h_1[i]-h_2[i])^2}{(h_1[i]+h_2[i])} \quad (1)$$

Selanjutnya, untuk menggabungkan invarian rotasi dalam pengukuran kemiripan, digunakan pengukuran jarak jauh Chi-square (SCSD) geser seperti yang diberikan di Persamaan berikut, di mana jumlah geser ditentukan oleh j dan $0 \le j \le nb - 1$.

$$dist(h1,h2)=\frac{1}{2}Min\left[\sum_{\forall i,\forall j}\frac{(h_1[i]-h_2[(i+j)\bmod n_b])^2}{(h_1[i]+h_2[(i+j)\bmod n_b])}\right] \quad (2)$$

Penelitian ini telah menunjukkan bagaimana titik akhir yang diekstraksi dari segmen garis lurus digital dapat digunakan secara efisien untuk menemukan konteks bentuk. Untuk melakukannya, telah menggunakan jajaran pecahan Farey yang sesuai dengan vektor arah untuk penemuan binning dan pencocokan yang kuat. Selanjutnya, menunjukkan bagaimana pendekatan ini dapat digunakan untuk analisis konteks bentuk bentuk karakter yang ditarik tangan. Ada ruang lingkup perpanjangan kerja lebih jauh ke arah pencocokan berbasis bentuk dan pengenalan karakter yang ditarik tangan.

Penelitian [2] mengusulkan penggabungan Shape Context dengan algoritma genetika kuantum (QGA) untuk menentukan metode pencocokan dan pengambilan bentuk baru. Metode pencocokan didasarkan pada penemuan korespondensi terbaik antara dua titik set. Metode yang diusulkan menggunakan QGA untuk meneemukan konfigurasi sampel poin terbaik untuk mencapai pencocokan terbaik antara dua bentuk.

Pencocokan bentuk didasarkan pada pengukuran kesamaan antara dua deskriptor bentuk, banyak bentuk representasi dan ukuran kemiripan bentuk dapat ditemukan dalam literatur. Bentuk deskriptor dapat dikelompokkan menjadi tiga kategori: metode berbasis kontur, berbasis wilayah dan titik set.

Metode berbasis kontur hanya menggunakan informasi bentuk kontur, yang menghadap bentuk informasi interior. Deskriptor berbasis wilayah memanfaatkan semua piksel bentuknya. Deskriptor momen, matriks lambung dan bentuk konveks adalah contoh metode berbasis wilayah. Teknik berbasis set point dapat didefinisikan sebagai seperangkat titik sampel yang diambil dari bentuk kontur. Bentuk konteks, skema pemungutan suara adalah metode yang paling populer dalam kategori ini. Dalam kebanyakan kasus poin dipilih secara acak dan memesan poin biasanya tidak diperlukan. Teknik ini didasarkan pada gagasan untuk menemukan korespondensi terbaik antara titik pada dua bentuk.

Pada penelitian ini, algoritma konteks bentuk kuantum diusulkan. Tujuan utama dari pekerjaan ini adalah untuk meningkatkan ketahanan dan kekuatan diskriminatif bentuk pencocokan dan pengambilan dengan menggunakan deskriptor konteks bentuk. Pendekatan QSC yang kami ajukan bermanfaat dari kekuatan algoritma genetika kuantum, yang hadir dalam bentuk superposisi kuantum, di mana setiap individu dapat mewakili tidak hanya satu solusi namun semua kemungkinan solusi untuk masalah tersebut. Pendekatan QSC mencakup semua konfigurasi titik sampel yang mungkin terjadi di ruang pencarian, dan juga menggunakan QGA untuk menemukan orientasi optimal memastikan hasil pencocokan bentuk yang lebih baik untuk bentuk yang diputar dan dibalik.

Penelitian [3] menggunakan deskriptor Shape Context pada jarak pengelompokkan yang tidak merata dan deskripsi bentuk fitur yang lebih luas, deskriptor ini memiliki target titik kontur yang mengatur inversi deformasi. Namun kode verifikasi yang membelit dan melekat memiliki banyak noise yang lebih serius, berbentuk yang sangat jelek, untuk mengatasi keterbatasan deskriptor diatas, maka meningkatkan algoritma baru berdasarkan bentuk relatif, titik pola yang cocok untuk mengidentifikasi kode.

CAPTCHA digunakan untuk membedakan perilaku manusia dan perilaku komputer otomatis, untuk mencegah registrasi dan cracking yang membahayakan. Manual Entry, dimana perlu memasukkan data satu per satu, sangat lambat dibandingkan dengan komputer dan hampir tidak berefek pada server. Latar belakang dan teknik pengenalan kode validasi didasarkan pada teknik pengenalan pola dan pengenalan gambar. Sebelum proses pengenalan gambar, dimana ada teknologi pra-pengolahan gambar mencakup gambar abu-abu, binerisasi, denoising, koreksi kemiringan, segmentasi karakter. Proses pengindeksan kode awal pre-processing. Teknologi pencocokan pola telah banyak diterapkan di bidang pengenalan karakter, pengenalan wajah, pengambilan gambar berbasis konten, dan pengawasan video cerdas. Ini adalah masalah mendasar untuk pemrosesan gambar dan pengenalan pola komputer.

Pencocokan pola didasarkan pada metode pengukuran berukuran mirip dengan bentuknya. Ada banyak metode pencocokan bentuk sesuai dengan metode klasifikasi yang berbeda, misalnya, sesuai kemampuannya dalam menghadapi titik transformasi menjadi terbagi sebagai berikut: pertama, dengan mencari invariant berbagai perubahan untuk menangani muatan bentuk, invarian ini adalah: Imitasi ditembak invarian, invariants serupa, invarian perspektif. Kedua, dengan menemukan fitur lokal antara target dan sampel untuk mendapatkan toleransi minimum yang sesuai untuk menangani deformasi yang lebih kompleks. Bila menggunakan bentuk atau pencocokan berbasis kontur dibagi antara dua jenis pendekatan berbasis struktur regional atau global ini. Proses pencocokan sampel dasar, tahap ekstraksi fitur dari proses diwakili oleh sebuah proses, tahap identifikasi pencocokan adalah salah satu proses identifikasi, dan estimasi parameter transformasi spasial

dan pelatihan iteratif adalah proses pembelajaran.

III. SHAPE CONTEXT

Shape Context [4]–[9] adalah sebuah fitur deskriptor yang digunakan untuk pengenalan objek. Shape Context merupakan suatu cara yang memungkinkan untuk mengukur kesamaan bentuk. Shape Context menggambarkan suatu bentuk dengan memperhatikan titik di dalam atau pada batas bentuk itu.

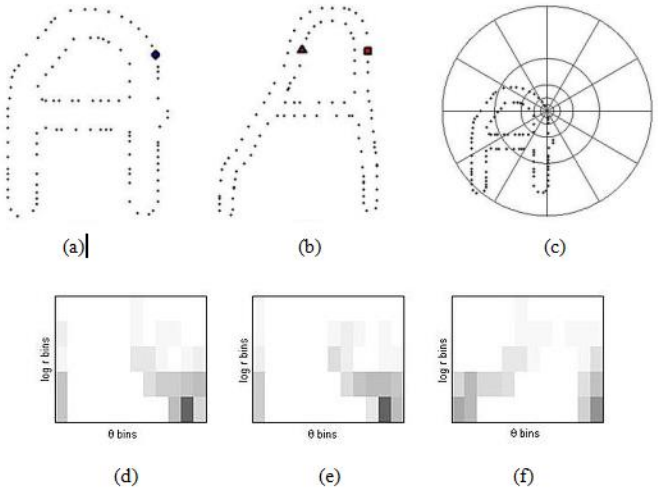
$$h_i(k) = \#\{Q \neq P_i : (Q - P_i) \in bin(k)\} \quad (3)$$

Persamaan diatas mendefinisikan Shape Context dari pi. Bin biasanya diambil dari sepasang ruang log-polar. Pada Gambar 3.2, ditunjukkan dimana terdapat perbedaan Shape Context antara dua versi huruf "A" yang berbeda.

Pada gambar 3.2 (a) dan (b) merupakan contoh gambar titik tepi untuk 2 bentuk yang berbeda. Pada gambar 3.2 (c) merupakan diagram bin log-polar yang digunakan untuk menghitung Shape Context. Pada gambar 3.2 (d) merupakan titik poin Shape Context yang ditandai dengan lingkaran pada gambar 3.2 (a). Pada gambar 3.2 (e) adalah titik poin Shape Context yang ditandai dengan diamond pada gambar 3.2 (b). Gambar 3.2 (f) merupakan titik poin Shape Context yang ditandai dengan segitiga. Shape Matching merupakan proses pencocokan objek dengan objek lain berdasarkan tingkat kemiripan bentuk dari objek tersebut.

Sistem yang menggunakan Shape Context untuk Shape Matching terdiri dari beberapa tahap antara lain

1. Secara acak pilih satu set poin yang berada ada tepi bentuk yang diketahui dan satu set poin lain pada tepi bentuk yang tidak diketahui.
2. Hitung Shape Context dari setiap poin pada proses pertama.
3. Cocokan setiap titik dari bentuk yang diketahui ke titik dari bentuk yang tidak diketahui.
4. Hitung "shape distance" antara setiap pasangan titik pada dua bentuk. Gunakan penjumlahan jarak dari Shape Context, jarak tampilan gambar dan energi lentur.
5. Untuk mengidentifikasi bentuk yang tidak diketahui, gunakan klasifikasi *nearest-neighbor* untuk membandingkan shape distance bentuk yang tidak diketahui dengan shape distance dengan bentuk yang diketahui



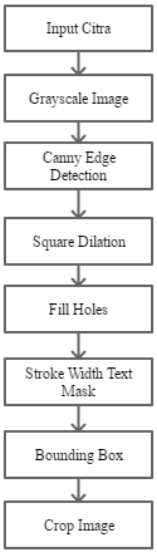
Gambar 3.1 Perbedaan Shape Context pada huruf A versi berbeda

IV. FAREY SHAPE CONTEXT

Pada Gambar 3.2 merupakan detil alur sistem yang menjelaskan proses untuk mendapatkan fitur Farey shape context untuk mendeteksi multiple object pada iklan di pinggir jalan yang melengkung sehingga didapatkan object setiap karakter tulisan berupa huruf atau angka untuk dilakukan proses selanjutnya yaitu pengenalan tulisan huruf atau angka.

Citra yang akan diproses dipilih terlebih dahulu. Inputan citra digunakan sebagai sumber utama untuk mendapatkan fitur shape context. Selanjutnya inputan citra diubah dalam level grayscale yaitu gambar hitam putih agar proses berikutnya dapat dilakukan dengan lebih mudah. Citra hasil Grayscale kemudian dilakukan proses deteksi tepi menggunakan Canny [10]. Algoritma Canny ini merupakan salah satu algoritma yang cukup populer dalam penggunaan pengolahan citra khususnya untuk mendeteksi tepi objek pada citra.

Citra hasil deteksi tepi menggunakan Canny tersebut diproses Square Dilation, dimana tepi-tepi pada objek-objek yang ada pada citra dilakukan penebalan tepi agar lebih mudah diproses pada tahap selanjutnya. Penebalan tepi ini juga dilakukan agar objek-objek pada citra dapat terdeteksi pada tahap selanjutnya, selain itu agar objek yang berukuran kecil juga tidak hilang dan dapat dikenali pada proses selanjutnya. Kemudian citra hasil Square Dilation dilakukan proses Fill Holes, dimana setiap objek yang ada pada citra hasil Square Dilation yang hanya terlihat tepi saja, akan diisi sehingga lebih terlihat sebagai kesatuan objek. Fill Holes ini dilakukan agar objek yang berlubang tersebut tidak dianggap sebagai 2 objek yang berbeda pada tahap selanjutnya.



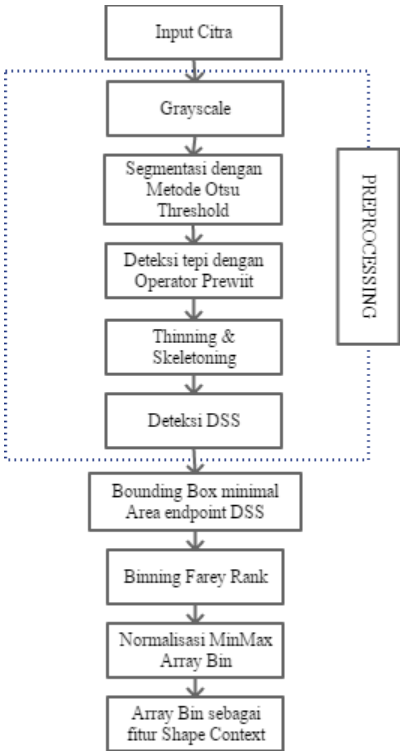
Gambar 3.2 Alur Sistem Deteksi Multiple Object

Citra hasil Fill Holes dilakukan proses Stroke Width Text Mask dengan tujuan untuk menghilangkan objek-objek yang bukan merupakan text yaitu huruf atau angka. Selanjutnya dilakukan proses Bounding Box, dimana setiap objek yang ada pada citra hasil Stroke Width Text Mask dibounding box untuk kemudian akan dilakukan pemotongan citra.

Lalu, citra yang telah di Bounding Box dilakukan proses Crop Image atau pemotongan citra. Pemotongan Citra ini dilakukan untuk mendapatkan citra dari setiap objek secara terpisah untuk melakukan proses pengenalan karakter setiap objek.

Gambar 3.3 Alur Sistem Farey Shape Context

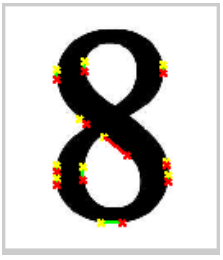
Citra yang akan diproses dipilih terlebih dahulu. Selanjutnya akan diproses Farey Shape Context per single object, seperti alur Farey Shape Context pada gambar 3.4. Input citra digunakan sebagai sumber utama untuk mendapatkan fitur shape context. Input citra diubah dalam level grayscale agar proses berikutnya dapat dilakukan dengan lebih mudah.



Gambar 3.4 Alur Sistem Farey Shape Context

Citra yang sudah digrayscale kemudian dilakukan proses segmentasi untuk memisahkan antara objek dengan latar belakang. Segmentasi menggunakan metode Otsu. Citra yang sudah disegmentasi kemudian dilakukan proses deteksi tepi dengan menggunakan operator Prewit sehingga didapatkan tepi-tepi dari objek citra yang kemudian digunakan untuk menemukan DSS (Digital Straight Line Segment). Deteksi tepi yang sudah didapatkan kemudian digunakan sebagai dasar untuk melakukan proses penipisan menggunakan metode Thinning dan Skeletoning. Deteksi tepi yang sudah didapatkan kemudian digunakan sebagai acuan untuk menemukan Digital Straight Line Segment dari citra. Garis lurus yang terdapat dalam citra akan dideteksi dalam tahap ini dengan tujuan agar endpoint dari DSS tersebut dijadikan sebagai titik fitur atau feature point untuk menemukan shape context dari citra.

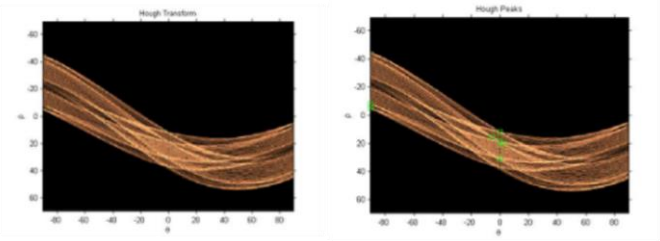
Berikut ini adalah hasil yang didapatkan setelah dilakukan proses deteksi DSS.



Gambar 3.5 Citra Hasil DSS

Pada gambar 3.5 merupakan hasil deteksi DSS dengan menggunakan metode Hough Transform. Hough Transform digunakan untuk mencari Hough Peaks. Hough Peaks

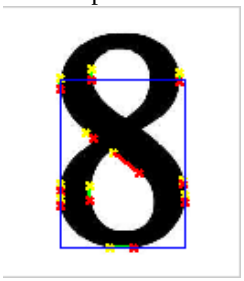
merupakan puncak dari transformasi Hough yang ditandai dengan kotak warna hijau pada grafik pada Gambar 3.6. Hough Peaks terdiri dari sumbu X dan Y yang mewakili rho dan theta, dimana rho merupakan resolusi jarak dari akumulator dengan satuan pixel dan theta merupakan resolusi sudut akumulator dengan satuan radian



Gambar 3.6 Hough Transform dan Hough Peaks

Garis lurus pada citra ditunjukkan dengan warna hijau sedangkan titik awal (startpoint) dan titik akhir (endpoint) masing-masing berwarna kuning dan merah. Total terdapat 11 garis lurus dengan 11 startpoint dan 11 endpoint yang didapatkan. Dalam kasus ini, yang diutamakan untuk digunakan sebagai feature point adalah endpoint dari DSS yang didapatkan. DSS endpoint yang didapatkan pada proses sebelumnya digunakan sebagai acuan untuk menemukan bounding box. Bounding box didapatkan berdasarkan minimum area dari endpoint yang didapatkan. Bounding box tersebut digunakan sebagai object boundary yang dimana setiap sudutnya merupakan reference point yang merupakan vektor menuju pada feature point yang didapatkan. Berikut ini adalah hasil bounding box endpoint yang didapatkan.

Pada gambar 3.7 merupakan hasil bounding box endpoint yang digunakan untuk menunjukkan shape context dari object. Dari gambar tersebut didapatkan reference point terdapat 4 titik dan feature point terdapat 11 titik. Sehingga vektor dari setiap reference point dan setiap feature point merepresentasikan pecahan dari Augmented Farey Sequence yang mencakup ruang 360º (setiap reference point mencakup 1 kuadran).



Gambar 3.7 Citra Hasil Bouding Box End Point

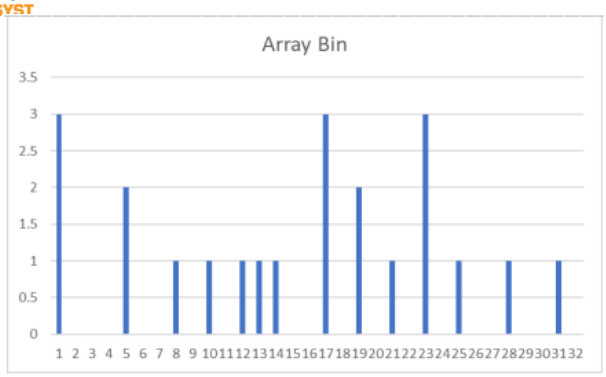
Untuk melakukan Binning Farey Rank, pertama-tama yang harus dilakukan adalah membentuk Augmented Farey Table (AFT). AFT dibentuk berdasarkan Augmented Farey Sequence yang merupakan pengembangan dari Farey Sequence. Beda Farey Sequence dan Augnented Farey Sequence adalah nilai pecahannya. Farey Sequence hanya meliputi pecahan dengan pembilang dan penyebut yang positif sedangkan Augmented Farey Sequence meliputi pecahan dengan pembilang dan penyebut positif serta negatif. Pada penelitian ini, AFT dibentuk dengan order 100 yang berarti pembilang dan penyebut akan bernilai $-100 \geq n \leq 100$.

Setelah didapatkan AFT maka proses selanjutnya adalah melakukan Binning Farey Rank. Proses pembentukan AFT bersifat independent artinya adalah tidak harus dilakukan bersamaan dengan proses binning tetapi harus dilakukan terlebih dahulu sebelum melakukan binning. Berikut ini adalah algoritma 1 yang digunakan untuk melakukan proses Binning Farey Rank.

Algoritma 1. Binning Farey Rank

```
1. Baca inputan citra.
2. Lakukan resize citra berdasarkan inputan yang
   diberikan.
3. Lakukan proses grayscale.
4. Lakukan Segmentasi dengan Otsu.
5. Lakukan deteksi tepi dengan operator Prewitt.
6. Lakukan Proses Thinning & Skeletoning.
7. Lakukan proses Hough Transform.
8. Cari Peak dari Hough transform yang didapatkan.
9. Cari Lines atau garis yang didapatkan oleh Hough
   Transform berdasarkan Peak yang didapatkan.
10.Lakukan perulangan sejumlah garis yang didapatkan.
11.  Temukan endpoint dari garis yang didapatkan.
12.  Simpan endpoint dalam array.
13.Lakukan bounding box berdasarkan minimum area dari
   endpoint  yang didapatkan.
14.Simpan Setiap titik sudut dari bounding box sebagai
   reference point.
15.Lakukan perulangan sejumlah garis yang didapatkan
   dengan index j.
  Lakukan perulangan sejumlah titik sudut bounding box
     dengan index k.
    Simpan selisih y dari endpoint ke-j dengan y dari
     reference point ke-k dalam variabel num_a.
    Simpan selisih x dari endpoint ke-j dengan x dari
     reference point ke-k dalam variabel num_b.
    Periksa apakah nilai mutlak dari num_a atau num_b
   lebih   besar dari nilai n.
    Jika ya maka
      Temukan closest fraction dari num_a sebagai
   pembilang dan num_b sebagai penyebut dari pecahan.
    Cari ranking pada AFT dengan num_a sebagai
   pembilang dan    num_b sebagai penyebut dan simpan
   dalam array features.
16.Insiialisasi variable n=32 dan total_fraction=12176.
17.Hitung width interval dari bin array yang ingin
   digunakan dengan w=total_fraction/n dan simpan dalam
   variabel w.
18.Lakukan perulangan sejumlah bin yang diberikan yaitu
   32 dengan index i.
  Isikan variabel interval ke (i,1) dengan w x(i-1).
  Isikan variabel interval ke (i,2) dengan minimum dari
     interval(i,1)+w-1 dengan total_fraction.
 Lakukan perulangan sejumlah features yang didapatkan
    dengan   index j.
   Periksa jika nilai features ke-j>=interval ke(i,1)
  dan   features ke-j<=interval ke (i,2).
    Jika ya maka
      Tambahkan bin ke-i dengan 1.
```

Pada paper ini jumlah bin yang digunakan adalah 32 untuk masing-masing objek yang diberikan. Sehingga array bin akan berjumlah 32. Array bin yang berjumlah 32 tersebut akan dibatasi dengan sejumlah interval untuk setiap binnya. Batasan interval berhubungan dengan nilai pecahan Augmented Farey yang didapatkan. Batasan interval tersebut juga memberikan pengaruh terhadap penambahan bobot pada setiap bin. Berikut ini adalah hasil array bin yang didapatkan setelah dilakukan proses binning.



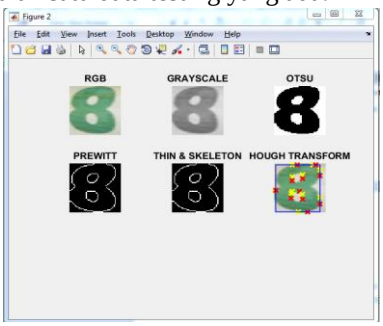
Gambar 3.7 Array Bin

Pada gambar 3.7 merupakan hasil array bin yang didapatkan setelah proses binning. Array yang didapatkan berjumlah 32 dengan nilai integer positif yang bervariasi. Setelah didapatkan array bin, proses selanjutnya adalah melakukan normalisasi dengan metode MinMax, agar memiliki bobot yang relevan dalam menunjukkan shape context dari object.

Pada kasus yang sebenarnya dalam pendeteksian text diketahui bahwa dapat terjadi object yang lebih dari satu dalam sebuah inputan citra. Jika object lebih dari satu maka sebelum melakukan perhitungan shape context, maka dilakukan proses deteksi multiple object terlebih dahulu.

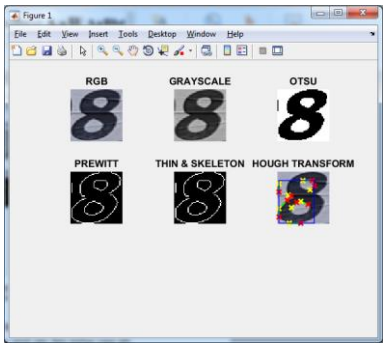
V. HASIL EKSPERIMEN DAN PENELITIAN

Pada bab ini akan dijelaskan hasil uji coba penelitian yang telah dilakukan pada beberapa gambar. Object Shape Context pada salah satu karakter angka 8, seperti pada Gambar 4.1, yang tertera pada salah satu data testing yang ada.



Gambar 4.1 Hasil Proses Object Shape Context

Angka tersebut diproses dengan RGB, grayscale, Otsu, Prewitt dan Thin & Skeleton hingga Hough Transform untuk mendapatkan startpoint dan endpoint angka tersebut.



Gambar 4.2 Hasil Proses Object Shape Context

Jika dibandingkan dengan Gambar 4.2, karakter angka tersebut sama yaitu 8, akan tetapi angka 8 pada Gambar 4.2 terlihat lebih miring ke kanan. Keduanya diproses dengan metode yang sama, akan tetapi hasil Hough Transform nya berbeda karena pada Gambar 4.2 terlihat lebih miring ke kanan sehingga startpoint dan endpointnya berbeda.

Pada penelitian ini digunakan 500 data iklan di pinggir jalan yang melengkung, dimana 70% digunakan sebagai data sampel. Dari 70% data training tersebut didapatkan ribuan karakter berupa huruf dan angka yang dijadikan data sample, seperti contoh data sample Angka 0 pada Gambar 4.3.

Gambar 4.3 Contoh Data Sample Angka 0

Setelah selesai proses training, maka akan dilakukan uji coba dengan meletakkan gambar pada direktori tertentu. Dengan dilakukan metode Grayscale, Canny Edge Detection, Square Dilation, Fill Holes, Stroke Width Text Mask, maka didapatkan pengenalan tulisan pada iklan pinggir jalan yang melengkung seperti pada Gambar 4.4

(a) (b) (c) (d)

(e) (f) (g)

Gambar 4.4 Proses Pengenalan Tulisan. (a) Gambar Asli, (b) Grayscale, (c) Canny Edge Detection, (d) Square Dilation, (e) Fill Holes, (f) Stroke Width Text Mask, (g) Detected Text

Dari contoh tersebut terlihat 12 huruf, 12 angka, sehingga total tulisan 24. Program berhasil mengenali 10 huruf, 12 angka, sehingga total tulisan yang dikenali yaitu 22. Tingkat akurasi Gambar 5.12 yaitu 91.66%.

Berdasarkan hasil uji coba penelitian yang dilakukan pada 500 Gambar dimana 30% sebagai data testing, maka hasil Farey Shape Context untuk mengenali tulisan berupa huruf dan angka pada iklan pinggir jalan yang melengkung mencapai akurasi benar 74.94% dan salah 25.06%.

VI. KESIMPULAN

Pada paper ini, telah ditunjukkan algoritma Farey Shape Context untuk mengenali tulisan pada iklan pinggir jalan yang melengkung. Berdasarkan hasil penelitian ini, penulis mengambil kesimpulan dari penelitian pengenalan tulisan pada iklan pinggir jalan yang melengkung, yaitu dari 500 data yang digunakan, dimana 30% data digunakan sebagai data testing, maka Pengenalan tulisan berupa huruf dan angka menggunakan Farey Shape Context dapat mencapai hasil akurasi kebenaran hingga 74.94%.

Dari hasil penelitian ini, penulis juga memberikan saran untuk penelitian selanjutnya bahwa Fitur Farey bergantung pada DSS endpoint, sehingga jika tidak ditemukan secara akurat dapat merusak hasil prediksi, maka disarankan untuk menambahkan fitur lain selain DSS endpoint yang mendukung DSS endpoint untuk dapat meningkatkan akurasi hasil prediksi.

Selain itu juga Fitur Farey menggunakan pecahan Augmented Farey Sequence maka jika terdapat sebuah kasus dimana pembilang dan penyebut lebih besar dari order n maka untuk mencari pecahan terdekat bisa jadi tidak ditemukan dan memberikan efek pada prediksi. Oleh karena itu, disarankan untuk menambahkan metode pencarian pecahan terdekat yang lebih baik sehingga meningkatkan akurasi prediksi.

DAFTAR PUSTAKA

[1] S. Pratihar and N. Begum, “Understanding shape context by analysis of Farey ranks,” in *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, 2016, pp. 580–585.

[2] K. M. Mezghiche, K. E. Melkemi, and S. Foufou, “Matching with quantum genetic algorithm and shape contexts,” in *2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)*, 2014, pp. 536–542.

[3] G. An and W. Yu, “Captcha recognition algorithm based on the relative shape context and point pattern matching,” in *2017 9th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, 2017, pp. 168–172.

[4] N. Bhuptani and B. Talati, “Variations in Shape Context Descriptor: A survey,” *Int. J. Comput. Appl.*, vol. 975, p. 8887, 2014.

[5] Y.-P. Lin and K.-W. Hsu, “Using Color Difference with Shape Context for Logo Recognition.,” *J. Softw.*, vol. 9, no. 8, pp. 2188–2193, 2014.

[6] S. Belongie, J. Malik, and J. Puzicha, “Shape context: A new descriptor for shape matching and object recognition,” *Adv. Neural Inf. Process. Syst.*, vol. 13, 2000.

[7] A. Thayananthan, B. Stenger, P. H. S. Torr, and R. Cipolla, “Shape context and chamfer matching in cluttered scenes,” in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, 2003, vol. 1, pp. I--I.

[8] J. Qi, L. Wenhui, L. Yi, and Y. YingTao, “An Efficient Object Recognition Method Based On Pyramid Match Kernel Using Shape Contexts,” in *2008 IEEE International Symposium on Knowledge Acquisition and Modeling Workshop*, 2008, pp. 18–21.

[9] S. G. Salve and K. C. Jondhale, “Shape matching and object recognition using shape contexts,” in *2010 3rd International Conference on Computer Science and Information Technology*, 2010, vol. 9, pp. 471–474.

[10] J. Canny, “A computational approach to edge detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 6, pp. 679–698, 1986.

Web Content Extractor Menggunakan Neural Network untuk Konten Artikel di Internet

Syabith Umar Ahdan, *Fakultas Sains dan Teknologi, Institut Sains dan Teknologi Terapan Surabaya*,
Joan Santoso, *Fakultas Sains dan Teknologi, Institut Sains dan Teknologi Terapan Surabaya*,
Hendrawan Armanto, *Fakultas Sains dan Teknologi, Institut Sains dan Teknologi Terapan Surabaya*.

***Abstrak*— Berkembangnya teknologi Javascript khususnya AJAX dan CSS membuat halaman web yang dulunya statis menjadi lebih dinamis dengan tampilan yang lebih menarik dan dipenuhi iklan dan rekomendasi artikel lain. Oleh karena itu, sulit untuk mengotomatisasi proses pengambilan konten artikel pada konteks ini. Penelitian ini dibuat untuk menyelesaikan masalah otomatisasi pengambilan konten artikel di Internet. Aplikasi web yang akan dibuat terbagi menjadi empat modul, yaitu web crawler, web extractor, content classifier dan web visualizer. Penelitian ini memiliki dua desain arsitektur. Arsitektur yang pertama adalah arsitektur saat training. Arsitektur yang kedua adalah arsitektur program jadi. Proses training menggunakan 200 URL halaman web dari lima website berbeda. Metode pengujian yang akan digunakan adalah 4-Fold Cross Validation, sehingga 75% dari blok teks akan menjadi data latihan dan 25% dari blok teks akan menjadi data pengujian. Program jadi berupa Web Visualizer yang mengolah JSON file berisi hubungan antara halaman web yang didapatkan dari web crawler sehingga dapat dipresentasikan dalam sebuah grafik. Kesimpulan dari penelitian ini adalah bahwa kombinasi Scrapy, Splash, Neural Network Classifier dan D3 bekerja sangat baik untuk automasi ekstraksi konten artikel website di Internet sekaligus memvisualisasi hubungan antar halaman web. Deep Feed Forward Neural Network (DFFNN) dapat melakukan klasifikasi multi-class konten judul, penulis, dan isi artikel dengan baik selama template halaman web sudah pernah dilatih sebelumnya. DFFNN juga dapat melakukan klasifikasi binari untuk halaman web secara umum dengan F1-score 62.87%, dua kali lebih baik dari SVM yang hanya 31.28%.**

***Kata Kunci*—Content Extractor, DBSCAN, Neural Network, Web Crawler, Web Visualization.**

I. PENDAHULUAN

Bagi peneliti literatur dan bahasa web, sangatlah penting untuk mendapatkan data penelitian sebanyak-banyaknya untuk kepentingan penelitiannya. Namun akan sangat memakan waktu jika data penelitian harus didapatkan dengan membuka halaman web secara manual satu per satu. Akan sangat membantu apabila ada suatu software yang dapat membantu peneliti literatur dan bahasa mengumpulkan data penelitian. Sebuah web parser biasanya digunakan untuk tujuan ini. Sangat mudah bagi sebuah parser untuk mengekstrak konten pada suatu halaman web, asalkan parser mampu mengenali struktur dari halaman web tersebut. Namun, dengan kecepatan pertambahan jumlah website yang ada, sangatlah sulit, bahkan mustahil, untuk menganalisis setiap website secara manual satu per satu. Struktur sebuah website juga dapat berubah sewaktu-waktu, sehingga akan mustahil juga untuk mengecek dan memperbaharui struktur website yang sudah ada.

Tujuan utama dari Penelitian ini adalah untuk mengotomatisasi proses pengambilan konten artikel di internet. Selain itu, dapat menyaring konten yang ada pada halaman web dengan hanya mengambil judul, penulis, dan konten pada halaman artikel. Tujuan lainnya adalah untuk memvisualisasi berbagai halaman website dengan mempresentasikan setiap halaman web sebagai sebuah node yang memiliki banyak koneksi ke halaman web yang lain.

II. TEORI PENUNJANG

Terdapat lima teori utama yang akan dibahas pada bagian ini. Kelima teori tersebut adalah Web Framework, Web Crawler, Headless Web Browser, Neural Network, dan Web Data Visualization.

A. *Web Framework*

Web Framework adalah koleksi *packages* dan *modules* yang memungkinkan developer untuk membuat aplikasi atau layanan web tanpa harus mengurusi detail *low-level* seperti manajemen protokol, soket, dan *process/thread*. Web Framework yang akan digunakan adalah Flask.

B. *Web Crawler*

Web Crawler, atau *spider* (laba-laba), adalah tipe bot yang biasanya dijalankan oleh *search engine* (mesin pencarian) seperti Google dan Bing. Tujuan web crawler adalah untuk mengindeks konten dari website yang ada di internet, mempelajari informasi apa yang ada pada website tersebut, sehingga website tersebut dapat muncul di hasil pencarian search engine saat informasi yang sama dibutuhkan pengguna search engine. Web Crawler Framework web crawler yang digunakan adalah Scrapy.

C. *Headless Web Browser*

Headless Web Browser adalah sebuah web browser yang tidak memiliki GUI (*Graphical User Interface)*. Headless

Syabith Umar Ahdan, Fakultas Sains dan Teknologi, Institut Sains dan Teknologi Terapan Surabaya, Surabaya, Jawa Timur, Indonesia (e-mail: syabith1@mhs.stts.edu)
Joan Santoso, Fakultas Sains dan Teknologi, Institut Sains dan Teknologi Terapan Surabaya, Surabaya, Jawa Timur, Indonesia (e-mail: joan@stts.edu)
Hendrawan Armanto, Fakultas Sains dan Teknologi, Institut Sains dan Teknologi Terapan Surabaya, Surabaya, Jawa Timur, Indonesia (e-mail: hendrawan@stts.edu)

web browser menyediakan kontrol otomatis pada halaman web di lingkungan yang mirip dengan web browser pada umumnya. Namun dijalankan menggunakan CLI (Command Line Interface) atau menggunakan komunikasi jaringan. Headless web browser sangat bermanfaat untuk pengujian halaman web karena dapat menerjemahkan dan mengerti HTML. Splash akan digunakan sebagai Headless Web Browser untuk Penelitian ini.

D. *Neural Network*

Neural Network adalah salah satu tipe *Machine Learning* (Pembelajaran Mesin) di mana modelnya dibuat berdasarkan cara kerja otak manusia. Melalui berbagai algoritma, model neural network dapat mengelompokkan dan mengklasifikasi sebuah dataset berdasarkan fitur-fitur yang diberikan sebagai input. Pola input yang dikenali neural network adalah pola angka yang dimuat dalam vector, dimana panjang vector sesuai dengan jumlah fitur. Perangkat lunak dan library yang akan digunakan untuk membangun neural network antara lain: Scikit Learn [1], Numpy , Keras [2], dan TensorFlow [3].

E. *Web Data Visualization*

Web Data Visualization atau visualisasi data web adalah teknik merepresentasikan grafik dari data dan informasi melalui media web. Representasi grafik dapat mempermudah pembaca untuk mengerti arti maupun implikasi dari data seperti tren, pola, *outliers*, maupun hubungan antar data. Komunikasi ini dicapai melalui pemetaan yang sistematis antara tanda grafik dengan nilai data yang dipresentasikan dalam visualisasi yang diciptakan. Pemetaan ini menetapkan bagaimana nilai data akan dipresentasikan secara visual. Pemetaan ini juga mendeterminasikan apa saja properti tanda grafik yang mengandung nilai data, seperti ukuran dan warna. Penelitian ini akan menggunakan D3 [4] sebagai tools pembuatan visualisasi data web.

III. Analisa Penelitian Sejenis

Terdapat dua penelitian yang akan digunakan sebagai rujukan. Penelitian pertama adalah *web content extractor through machine learning* yang ditulis oleh Ziyan Zhou et al [5]. Penelitian yang kedua adalah WebSPHINXdari Computer Science of Carnegie Mellon University.

A. *Web Content Extractor Through Machine Learning*

Paper ini dipilih karena kebanyakan paper web ekstraktor berupaya untuk mengekstrak data yang terstruktur dan jumlahnya ada beberapa dalam satu halaman. Hal ini dicapai dengan menggunakan model classifier SVM atau Support Vector Machine [6]. Sedangkan tujuan dari paper tersebut adalah untuk mengekstrak konten web yang hanya dimuat sekali oleh sebuah halaman web, seperti konten artikel, berita, maupun cerita. Tujuan paper tersebut selaras dengan tujuan dari Penelitian ini.

Paper ini terdiri dari enam bagian. Bagian-bagian tersebut antara lain pengumpulan data, ekstraksi blok teks dan CSS, klasterisasi, pelabelan klaster, SVM dan cross validation, dan pemilihan fitur.

B. *WebSPHINX*

WebSPHINX adalah sebuah library Java Class dan lingkungan development yang interaktif untuk web crawler. WebSPHINX ditujukan untuk pengguna web lanjutan dan programmer Java yang ingin melakukan crawling pada sebagian kecil dari sebuah website secara otomatis. WebSPHINX dirilis secara open source, di bawah Apache-style license.

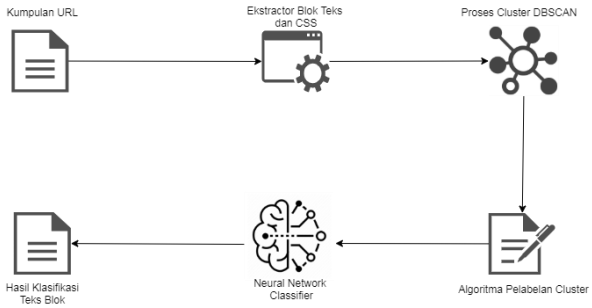
Bagian WebSHINX yang akan dirujuk adalah bagian Crawler Workbench. Crawler Workbench adalah sebuah graphical user interface yang memungkinkan penggunanya untuk mengkonfigurasi dan mengkontrol web crawler yang terkostumisasi. Dengan Crawler Workbench, pengguna dapat memvisualisasi kumpulan halaman web dalam sebuah grafik, menyimpan halaman web ke tempat penyimpanan lokal di komputer untuk browsing secara offline, menggabungkan beberapa halaman menjadi satu dokumen untuk dibaca atau dicetak, mengekstrak teks tertentu dengan mencocokkan pola dari kumpulan halaman web, dan membuat crawler yang dapat dikostumisasi menggunakan Java atau Javascript agar dapat memproses halaman web sesuai keinginan.

IV. Arsitektur Sistem

Penelitian ini memiliki dua desain arsitektur. Arsitektur yang pertama adalah arsitektur saat training. Arsitektur yang kedua adalah arsitektur program jadi.

A. *Arsitektur Saat Training*

Arsitektur saat training adalah arsitektur sistem yang digunakaan saat melatih model neural network. Model neural network sering digunakan untuk kebutuhan klasifikasi teks [7], [8]. Desain arsitektur saat training model neural network dari Penelitian ini dapat dilihat pada Gambar 1.



Gambar. 1. Desain Arsitektur Saat Training

Proses training dimulai dengan pengumpulan 200 URL halaman web dari lima website yaitu detik.com, kompas.com, iwanbanaran.com, teknojurnal.com, dan tipspintar.com. Kira-kira seperempat diantara 200 URL tersebut adalah halaman web yang tidak berartikel.

Ekstraktor blok teks dan CSS akan membuka tiap halaman web dari daftar URL seperti halnya sebuah web browser. Ekstraktor blok teks dan CSS kemudian mengekstrak setiap elemen teks blok beserta tag path dan properti CSS nya dari setiap halaman web.

Proses Cluster DBSCAN akan memproses kumpulan file json yang diproduksi oleh Ekstraktor blok teks dan CSS untuk mengumpulkan teks blok yang sejenis di satu klaster yang sama menggunakan algoritma klasterisasi DBSCAN [9].

Hasil dari klasterisasi ini dapat berupa belasan hingga ratusan klaster tergantung pada karakteristik layout website.

Algoritma pelabelan cluster kemudian menggunakan hasil dari proses cluster DBSCAN untuk melabeli klaster terbaik berdasarkan nilai tertinggi. Klaster ini dihitung menggunakan algoritma LCS yang memanfaatkan tag meta dari sebuah halaman web dan algoritma buatan Ziyan Zhou untuk mengatasi pemberian nilai yang tidak akurat pada blok teks berisi komentar. Blok teks yang ada dalam klaster dengan nilai terbaik tersebut kemudian dilabeli sebagai konten, sedangkan blok teks di dalam klaster lainnya dilabeli sebagai non konten. Klaster yang berisi judul dan penulis kemudian dilabeli secara manual.

Blok teks yang sudah dilabeli kemudian dijadikan sebagai dataset untuk melatih dan menguji model neural network. Kombinasi dari properti CSS dan tag path dari setiap blok teks dijadikan sebagai fitur atau input untuk model neural network yang akan dilatih. Metode pengujian yang akan digunakan adalah 4-Fold Cross Validation, sehingga 75% dari blok teks akan menjadi data latihan dan 25% dari blok teks akan menjadi data pengujian.

Selanjutnya akan dijelaskan tentang arsitektur dari model neural network yang akan dilatih. Matriks input untuk model neural network dapat dinotasikan sebagai $X \in \mathbb{R}^{188_x \times 35177}$. X adalah variabel matriks input berisi bilangan riil dengan ukuran 188 x 35177. Sedangkan sebuah sampel dataset ke-i dapat dinotasikan sebagai $x^{(i)} \in \mathbb{R}^{188_x}$, dimana x adalah vektor input berisi bilangan riil dengan ukuran 188.

Metode pengujian yang akan digunakan adalah 4-Fold Cross Validation, sehingga 75% dari blok teks akan menjadi data latihan dan 25% dari blok teks akan menjadi data pengujian. Pembagian data latihan dan data pengujian dilakukan per website, sebelum akhirnya dataset per website disatukan.

Untuk label kelas, dilakukan binarisasi, dimana cara kerjanya hampir sama dengan One Hot Encoding. Label yang semula berupa integer dengan kemungkinan nilai angka 0-4 diubah menjadi kumpulan vektor binari dengan kemungkinan angka 0 sampai 1. Panjang dari vektor binari ini adalah sebesar jumlah kelas klasifikasi. Penelitian ini memiliki empat kelas klasifikasi, sehingga panjang vektor binari adalah empat.

Matriks label dari model neural network pada Penelitian ini dapat dinotasikan sebagai $Y \in \mathbb{R}^{4_y \times 35177}$. Y adalah variabel matriks label berisi bilangan riil dengan ukuran 4 x 35177. Sedangkan sebuah label untuk dataset ke-i dapat dinotasikan sebagai $y^{(i)} \in \mathbb{R}^{4_y}$, dimana y adalah vektor label berisi bilangan riil dengan ukuran empat. Setiap indices dari vektor mencerminkan label dari dataset tersebut, dimana index pertama adalah non konten, index kedua adalah judul, index ketiga adalah penulis, dan index keempat adalah konten.

Matriks weight yang pertama untuk model neural network dapat dinotasikan sebagai $W^{[1]} \in \mathbb{R}^{96 \times 188}$. $W^{[1]}$ adalah variabel matriks weight berisi bilangan riil dengan ukuran 96 x 188. Matriks weight yang kedua untuk model neural network dapat dinotasikan sebagai $W^{[2]} \in \mathbb{R}^{20 \times 96}$. $W^{[2]}$ adalah variabel matriks weight berisi bilangan riil dengan ukuran 20 x 96. Matriks weight yang ketiga untuk model neural network dapat dinotasikan sebagai $W^{[3]} \in \mathbb{R}^{4 \times 20}$. $W^{[3]}$ adalah variabel matriks weight berisi bilangan riil dengan ukuran 4 x 20. Rumus fungsi aktivasi pada hidden layer pertama dapat dilihat pada formula 1.

$$a^{[1]} = \mathrm{ReLu}^{[1]}(W^{[1]}x^{(i)} + b_1) \quad (1)$$

Fungsi aktivasi yang digunakan pada hidden layer pertama adalah fungsi ReLu. Input dari fungsi ReLu adalah jumlah dari hasil perkalian matriks $W^{[1]}$ dengan vektor sampel x ke-i ditambah dengan nilai bias b_1. Rumus fungsi aktivasi pada hidden layer kedua dapat dilihat pada formula 2.

$$a^{[2]} = \mathrm{ReLu}^{[2]}(W^{[2]}h_1^{(i)} + b_2) \quad (2)$$

Fungsi aktivasi yang digunakan pada hidden layer kedua adalah fungsi ReLu. Input dari fungsi ReLu adalah jumlah dari hasil perkalian matriks $W^{[2]}$ dengan vektor sampel h_1 ke-i dari hidden layer sebelumnya ditambah dengan nilai bias b_2. Rumus fungsi aktivasi pada output layer dapat dilihat pada formula 3 di bawah ini.

$$\hat{y}^{(i)} = \mathrm{Softmax}(W^{[3]}h_2^{(i)} + b_3) \quad (3)$$

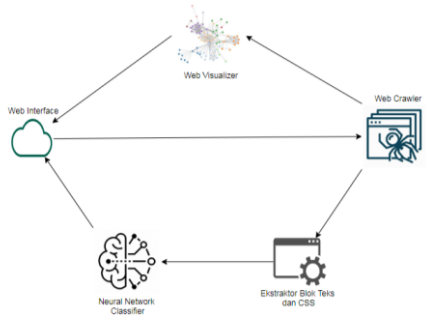
Fungsi aktivasi yang digunakan pada output layer adalah fungsi Softmax. Inputnya adalah jumlah dari hasil perkalian matriks $W^{[3]}$ dengan vektor sampel h_2 ke-i dari hidden layer sebelumnya ditambah dengan nilai bias b_3.

Optimizer yang digunakan pada model neural network adalah optimizer adam [10]. Adam merupakan singkatan dari Adaptive Momen Estimation. Adam mengkombinasikan properti terbaik dari AdaGrad dan RMSProp untuk menangani gradient yang jarang dan masalah noise. Adam sangatlah mudah dikonfigurasi dan konfigurasi defaultnya biasanya bekerja baik untuk kebanyakan masalah. Adam memiliki beberapa manfaat lain yaitu komputasinya yang efisien dan kebutuhan kapasitas memory nya yang kecil.

Loss function yang digunakan adalah categorical cross entropy. Categorical cross entropy dipakai karena jumlah kelas untuk klasifikasi adalah empat kelas, yaitu konten, judul, penulis, dan non konten. Karena terdapat ketimpangan yang tinggi antara blok teks yang berlabel non konten dengan blok teks yang berlabel konten, judul, dan penulis, maka fitur class weight saat latihan akan digunakan. Dengan fitur class weight, model neural network dapat disetting agar menyesuaikan nilai loss function setiap kelas sesuai dengan nilai class weight. Nilai dari class weight akan ditentukan sesuai dengan jumlah dataset pada setiap kelas. Dalam kata lain, kelas konten, judul, dan penulis, yang jumlah sampel datanya jauh lebih sedikit ketimbang sampel data non konten, akan memiliki nilai class weight yang jauh lebih besar ketimbang kelas non konten. Dengan ini diharapkan permasalahan dataset yang tidak seimbang dapat diatasi.

B. Arsitektur Program Jadi

Arsitektur program jadi digunakan untuk website yang akan digunakan oleh pengguna. Desain arsitektur program jadi dapat dilihat pada Gambar 2.



Gambar. 2. Desain Arsitektur Program Jadi

Sistem dimulai dari beroperasinya web interface yang bertatap muka langsung dengan pengguna website. Dari sini, pengguna website akan memberi input berupa base URL atau URL pertama yang akan dijadikan target crawling, depth atau jumlah kedalaman crawling dari base URL, dan total maksimal jumlah halaman web. Pengguna juga dapat memilih apakah crawler hanya akan menargetkan domain yang sama dari base URL, atau akan menargetkan semua out link yang ditemui.

Setelah input dan konfigurasi disubmit, maka web crawler akan bekerja sesuai nilai dari input dan konfigurasi. Web Crawler kemudian akan mengeluarkan output berupa daftar URL yang telah dicrawl dengan jumlah sesuai dengan nilai input dan juga file JSON yang menyimpan data tentang hubungan antar link URL.

Ekstraktor blok teks dan CSS pada program jadi adalah modul yang sama persis seperti yang digunakan pada saat training model neural network. Ekstraktor blok teks dan CSS akan mengekstrak blok teks dari daftar URL.

Neural Network Classifier kemudian memproses lebih lanjut kumpulan file JSON dari ekstraktor agar bisa dijadikan sebagai input untuk classifier. Hal ini dilakukan sebelum akhirnya diklasifikasi apakah sebuah data set termasuk dalam konten, penulis, judul, atau non konten. Data hasil dari klasifikasi ini kemudian diberikan kepada web interface untuk selanjutnya dipresentasikan.

Web Visualizer mengolah JSON file berisi hubungan antara halaman web yang didapatkan dari web crawler sehingga dapat dipresentasikan dalam sebuah grafik. Grafik tersebut berisi kumpulan node yang melambangkan sebuah halaman web dengan garis penghubung antar node yang melambangkan adanya link penghubung diantara keduanya. Grafik ini kemudian diberikan kepada web interface untuk selanjutnya ditunjukkan pada pengguna.

V. UJI COBA

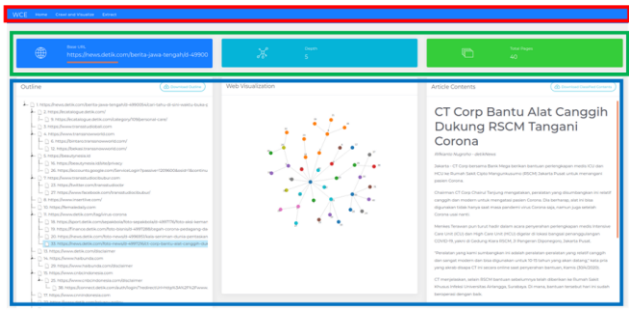
Pada bab uji coba sistem ini akan dijelaskan lebih detail mengenai hasil uji coba dan penjelasan detailnya dari website yang akan dibuat oleh Penelitian ini. Bab ini akan dibagi menjadi tujuh bagian, yaitu Web Interface, Web Crawler, Ekstraktor Blok Teks dan CSS, Proses Cluster DBSCAN, Algoritma Pelabelan Cluster, Neural Network Classifier, dan Web Data Visualizer.

A. Web Interface

Pada halaman utama, pengguna dikenalkan dengan website Penelitian ini beserta deskripsinya dan dapat langsung menggunakannya dengan mengisi form yang ada pada halaman yang sama. Form ini berisi data yang akan dijadikan input untuk modul web crawler.

Pada halaman utama website, terdapat tiga komponen utama, yaitu navbar dalam kotak merah, penjelasan tentang fungsi website dalam kotak hijau, dan form input dalam kotak biru. Navbar berhasil ditampilkan pada area atas halaman website di setiap halaman web. Setiap link pada Navbar berhasil membawa membawa pengguna kepada halamannya masing-masing, begitu juga dengan link pada logo website. Penjelasan tentang fungsi website berisi deskripsi tentang website dan penjelasan tentang cara penggunaannya. Form input berisi input base URL, depth, dan total maksimal jumlah halaman web. Form input juga berisi switch konfigurasi untuk mengkonfigurasi web crawler agar hanya menargetkan link dari domain yang sama dengan starting atau base URL. Dan terakhir, form input memiliki button submit yang mengirimkan data form ke proses selanjutnya.

Pada halaman hasil utama, pengguna dapat melihat hasil dari klasifikasi dan visualisasi website. Pengguna juga dapat melihat rekap dari input yang dimasukkan sebelumnya. Tampilan halaman hasil utama website dapat dilihat pada Gambar 3 di bawah ini.

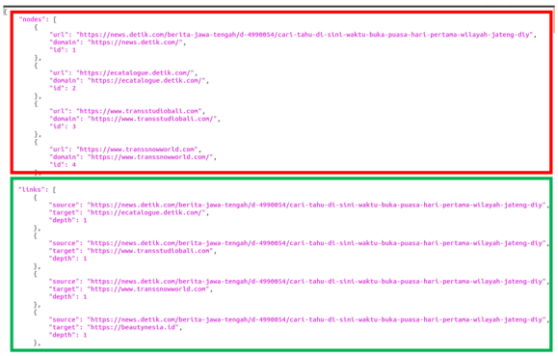


Gambar. 3. Halaman Hasil Utama Website

Pada halaman hasil utama website, terdapat tiga komponen utama, yaitu navbar dalam kotak merah di atas, rekap dari input yang dimasukkan sebelumnya dalam kotak hijau di tengah, dan hasil dari klasifikasi dan visualisasi dalam kotak biru di bawah. Navbar sudah dijelaskan pada paragraf sebelumnya. Rekap dari input yang dimasukkan sebelumnya berisi input base URL, depth, dan total maksimal jumlah halaman web. Hasil dari klasifikasi dan visualisasi berisi outline tree dan grafik hasil web crawler serta hasil klasifikasi dari neural network classifier.

B. Web Crawler

Web Crawler berhasil melakukan crawling sesuai dengan parameter dan konfigurasi yang diberikan. Web Crawler juga berhasil menghasilkan file teks daftar URL dan file JSON yang berisi daftar nodes dan daftar link antar nodes. Tampilan file JSON dapat dilihat pada Gambar 4.



Gambar. 4. File Daftar Nodes dan Link

Gambar 4 menunjukkan empat nodes pertama dalam kotak merah dan empat links pertama dalam kotak hijau dari 40 nodes dan links yang telah dihasilkan oleh web crawler. Base URL dapat dilihat pada URL node pertama pada file JSON yang dihasilkan. Setiap node yang dihasilkan memiliki tiga atribut, yaitu URL, domain, dan id. Sedangkan setiap link yang dihasilkan memiliki tiga atribut juga, yaitu source, target, dan depth. Hasil file daftar nodes yang dihasilkan membuktikan bahwa web crawler dapat menghasilkan daftar nodes sesuai dengan data dan atribut yang diinginkan. Hasil file ini juga membuktikan bahwa web crawler hanya akan memasukkan nodes yang memiliki source URL yang sudah ada pada daftar nodes saat itu. Sehingga setiap node pada daftar dapat dipastikan memiliki link dengan node lain. Dan terakhir, file daftar nodes yang dihasilkan berhasil membatasi jumlah depth dan jumlah total halaman web yang dihasilkan sesuai dengan input.

C. Ekstraktor Blok Teks dan CSS

Ekstraktor Blok teks berhasil melakukan ekstraksi blok teks dan CSS sesuai daftar URL yang diberikan. Setiap blok teks memiliki tujuh atribut, yaitu bound, computed, element, html, text, path, dan selector. Bound adalah lebar dan tinggi serta koordinat dari blok teks. Computed adalah properti CSS blok teks. Element adalah elemen tag HTML beserta atribut id dan class nya, html adalah isi raw HTML blok teks. Text adalah daftar isi teks pada blok teks. Path adalah tag path dari elemen sebelum tag body sampai pada blok teks tersebut. Selector adalah daftar CSS selector pada setiap elemen yang ada pada tag path. Blok teks di atas merupakan konten AJAX dari halaman web. Oleh karena itu, hasil uji coba ini membuktikan pula bahwa ekstraktor blok teks dan CSS dapat menangani konten AJAX.

D. Proses Cluster DBSCAN

Proses Cluster DBSCAN berhasil mengklasiterisasi kumpulan blok teks yang diberikan. Rincian hasil jumlah klaster dapat dilihat pada Tabel I di bawah ini.

TABEL I
JUMLAH KLASTER WEBSITE LATIHAN

Website	Jumlah Klaster
Detik.com	1116
Iwanbanaran.com	507
Kompas.com	859
Teknojurnal.com	141
Tipspintar.com	215
Total	200

Website dengan hasil klasterisasi yang terbaik adalah website dengan jumlah klaster yang paling sedikit. Maksud dari hasil klasterisasi terbaik disini adalah hasil klaster yang menyatukan semua blok teks yang penampilannya mirip kedalam klaster yang sama. Dalam hal ini, Teknojurnal.com dan Tipspintar mendapatkan posisi pertama dan kedua secara berurutan sebagai website dengan klaster terbaik. Namun sayangnya, hasil klaster dari kedua website tersebut masih kurang ideal. Pada Teknojurnal, blok teks yang merupakan isi artikel masih terbagi menjadi dua klaster. Pada Tipspintar.com, blok teks berisi artikel terbagi ke dalam empat klaster. Hal ini dikarenakan tidak semua blok teks yang berisi artikel memiliki properti CSS dan tag path yang 100% sama. Terdapat sedikit perbedaan yang menyebabkan blok teks berisi artikel terpisah ke beberapa klaster yang berbeda.

Masalah yang sama juga dimiliki oleh ketiga website yang lain. Detik.com dan Kompas.com memiliki hasil klasterisasi yang terburuk karena selain masalah properti CSS yang agak berbeda, kedua website tersebut memiliki template yang berbeda-beda pula. Template yang dipakai tergantung pada sub kategori artikel yang tertulis. Sehingga, blok teks yang berisi artikel terbagi menjadi beberapa klaster berdasarkan template pada halaman tersebut.

E. Algoritma Pelabelan Cluster

Telah ditentukan sebelumnya bahwa klasterisasi dari proses cluster DBSCAN kurang ideal. Oleh karena itu, diputuskan bahwa blok teks akan dilabeli secara manual untuk memastikan integrasi dan keakuratan dataset yang akan digunakan untuk melatih model neural network.

Meskipun begitu, algoritma pelabelan cluster akan tetap dicoba. Hal ini dilakukan untuk melihat apakah algoritma ini dapat digunakan pada website berbahasa Indonesia atau tidak. Dengan catatan bahwa kumpulan klaster yang akan diperiksa adalah kumpulan klaster yang baik. Dalam kata lain, algoritma pelabelan cluster ini hanya akan digunakan pada dataset website Iwanbanaran.com, Teknojurnal.com, dan Tipspintar.com yang hanya memiliki satu template.

Hasil yang didapatkan dari algoritma pelabelan cluster pada website Iwanbanaran.com, Teknojurnal.com, dan Tipspintar.com sedikit beragam. Pada Iwanbanaran.com, klaster dengan skor tertinggi adalah klaster dengan blok teks yang berisi username dan tanggal komentar. Setelah ditelaah lebih jauh, ternyata pada salah satu artikel terdapat tag meta yang berisi bulan dan tahun. Pada kolom komentar, terdapat teks bulan dan tahun pada setiap komentarnya juga. Hal ini menjelaskan kenapa algoritma LCS memberi nilai tertinggi pada klaster berisi komentar. Enam klaster terbaik jatuh pada blok teks komentar. Klaster dengan blok teks berisi konten jatuh pada posisi ketujuh. Pada Teknojurnal.com, klaster dengan skor tertinggi pertama dan kedua adalah klaster dengan blok teks yang berisi konten artikel. Pada Tipspintar.com, klaster dengan skor tertinggi pertama dan kedua adalah klaster dengan blok teks yang berisi konten artikel juga. Hasil dari pelabelan manual blok teks pada setiap website dapat dilihat pada Tabel II.

TABEL II
HASIL PELABELAN BLOK TEKS

Website	Non Konten	Judul	Penulis	Konten
Detik.com	4551	25	25	219
Iwanbanaran.com	9583	30	30	282
Kompas.com	9858	28	97	333
Teknojurnal.com	1645	29	29	383
Tipspintar.com	5724	30	58	2218
Covid19.go.id	246	3	0	32
Turnbackhoax.id	787	3	3	78
Total	32394	148	242	3545

F. Neural Network Classifier

Metode evaluasi model neural network yang akan digunakan adalah metode Precision, Recall, dan F-1 Score. Hasil uji coba yang akan ditampilkan adalah hasil yang menggunakan dataset dari website latihan dan dataset dari website baru yang belum pernah dilihat sebelumnya. Performa dari model neural network dasar dengan arsitektur dan konfigurasi yang sudah dideskripsikan sebelumnya dapat dilihat pada Tabel III di bawah ini.

TABEL III
PERFORMA MODEL DASAR

	Website Latihan			Website Baru		
Label	P	R	F-1	P	R	F-1
Non-Konten	99.66%	99.73%	99.69%	96.26%	87.12%	91.46%
Judul	97.06%	97.06%	97.06%	100%	50%	66.67%
Penulis	96.97%	91.43%	94.12%	0%	0%	0%
Konten	97.89%	97.67%	97.78%	37.75%	70%	49.04%

Model dasar neural network ideal digunakan pada website yang sudah dilatih sebelumnya. Namun model ini tidak ideal digunakan untuk website baru yang belum pernah dilatih.

Model selanjutnya adalah model dasar yang dimodifikasi hanya menggunakan satu hidden layer dengan jumlah node 96. Dalam kata lain, hidden layer kedua dengan jumlah node 20 tidak digunakan pada model ini. Performa variasi model dengan satu hidden layer dapat dilihat pada Tabel IV.

TABEL IV
PERFORMA MODEL SATU HIDDEN LAYER

	Website Latihan			Website Baru		
Label	P	R	F-1	P	R	F-1
Non-Konten	99.72%	99.69%	99.70%	90.71%	93.61%	92.14%
Judul	100%	97.06%	98.51%	0%	0%	0%
Penulis	97.06%	94.29%	95.65%	0%	0%	0%
Konten	97.45%	98.02%	97.73%	25.4%	14.55%	18.5%

Pada website yang sudah dilatih, model neural network dengan satu hidden layer memiliki performa yang hampir sama, bahkan sedikit lebih baik, dibandingkan dengan model dasar yang memiliki dua hidden layer. Namun, performa model dengan satu hidden layer jauh lebih buruk pada website yang belum pernah dilatih sebelumnya. Hal ini terjadi karena tingkat abstraksi dan generalisasi model dengan satu hidden layer jauh lebih rendah dibandingkan dengan model dengan dua atau lebih hidden layer.

Model selanjutnya adalah model dasar yang dimodifikasi dengan dua layer Dropout dengan dropout rate sebesar 0.5. Layer dropout pertama diletakkan diantara hidden layer pertama dan kedua. Layer dropout kedua diletakkan diantara hidden layer kedua dan output layer. Performa variasi model dengan dropout layer dapat dilihat pada Tabel V.

TABEL V
PERFORMA MODEL DROPOUT LAYER

	Website Latihan			Website Baru		
Label	P	R	F-1	P	R	F-1
Non-Konten	99.66%	99.61%	99.64%	95.36%	87.71%	91.38%
Judul	100%	88.24%	93.75%	0%	0%	0%
Penulis	88.57%	88.57%	88.57%	0%	0%	0%
Konten	96.99%	97.78%	97.39%	51.51%	30.91%	38.64%

Pada website yang sudah dilatih, model neural network dengan dua dropout layer memiliki performa yang sepadan dengan model neural network tanpa dropout layer. Namun hal ini berlaku hanya pada label kelas yang memiliki banyak sampel. Performa klasifikasi label kelas dengan sampel sedikit seperti judul dan penulis justru sedikit menurun. Sedangkan pada website baru yang belum pernah dilatih sebelumnya, performa model ini cenderung menurun.

Model selanjutnya adalah model dasar yang dilatih hanya menggunakan properti CSS sebagai fitur. Performa variasi model ini dapat dilihat pada Tabel VI.

TABEL VI
PERFORMA MODEL PROPERTI CSS

	Website Latihan			Website Baru		
Label	P	R	F-1	P	R	F-1
Non-Konten	99.71%	99.45%	99.58%	98.43%	85.09%	91.28%
Judul	100%	94.12%	96.97%	0%	0%	0%
Penulis	96.92%	90%	93.33%	0%	0%	0%
Konten	95.25%	98.37%	96.79%	40%	92.73%	55.89%

Pada website yang sudah dilatih, model neural network yang hanya menggunakan properti CSS sebagai fitur memiliki performa yang sepadan dengan model neural network dasar dengan tag path dan properti CSS sebagai fitur. Pada website yang belum pernah dilatih sebelumnya, performa identifikasi konten lebih baik jika hanya mengandalkan properti CSS. Hal ini mungkin dikarenakan karakteristik properti CSS yang cenderung lebih objektif dan umum pada berbagai website jika dibandingkan dengan tag path. Namun penggunaan tag path juga penting untuk mengidentifikasi suatu judul pada website secara umum. Hal ini dikarenakan kebanyakan blok teks judul pada suatu website ditandai dengan tag heading ketimbang tag paragraf.

Model selanjutnya adalah model dasar yang dilatih hanya menggunakan dataset website yang sama saja. Sehingga terdapat lima total model yang dilatih khusus untuk masing-masing Detik.com, Iwanbanaran.com, Kompas.com, Teknojurnal.com, dan Tipspintar.com. Performa variasi model per website ini dapat dilihat pada Tabel VII.

TABEL VII
PERFORMA F-1 SCORE MODEL PER WEBSITE

Website	Non-Konten	Judul	Penulis	Konten
Detik.com	99.99%	100%	88.89%	100%
Iwanbanaran.com	99.98%	100%	100%	99.13%
Kompas.com	99.61%	95.24%	90%	92.94%
Teknojurnal.com	99.88%	100%	100%	99.43%
Tipspintar.com	99.19%	100%	100%	98%

Model neural network yang dilatih hanya dengan

menggunakan dataset suatu website saja dapat mengidentifikasi judul, penulis, dan konten dengan hampir sempurna. Namun model neural network yang sudah terlatih ternyata sedikit sensitif terhadap noise sehingga mencegah performa model untuk mencapai nilai sempurna. Masalah ini tidak dimiliki oleh model Support Vector Machine yang dijadikan sebagai rujukan.

Selanjutnya model klasifikasi binari akan di uji coba. Model binary class yang pertama adalah model binary class yang hampir sama dengan model dasar, namun jumlah node pada output layernya hanya satu. Node ini merepresentasikan apakah suatu blok teks merupakan konten atau tidak. Loss Function yang digunakan juga diganti dengan Binary Crossentropy. Model ini selanjutnya akan direferensikan sebagai model dasar klasifikasi binari. Model kedua adalah model dasar klasifikasi binari yang datasetnya hanya menggunakan properti CSS sebagai fitur. Model ketiga adalah model dasar klasifikasi binari yang tidak menggunakan class weight. Model keempat adalah model dasar klasifikasi binari yang menggunakan dua dropout layer dengan dropout rate 0.2. Performa setiap model klasifikasi binari diatas dapat dilihat pada Tabel VIII.

TABEL VIII
PERFORMA MODEL KLASIFIKASI BINARI

	Website Latihan			Website Baru		
Website	**P**	**R**	**F-1**	**P**	**R**	**F-1**
Model 1	96.82%	98.23%	97.52%	16.19%	14.29%	15.18%
Model 2	95.51%	97.50%	96.5%	11.46%	9.24%	10.23%
Model 3	97.40%	97.61%	97.51%	57.24%	69.75%	62.87%
Model 4	97.5%	97.4%	97.45%	21.05%	26.89%	23.62%

Performa semua variasi model neural network pada website yang sudah dilatih sedikit lebih baik ketimbang SVM. Performa model neural network yang dilatih tanpa menggunakan class weight memberikan performa terbaik. Performa F-1 score yang dihasilkan bahkan dua kali lebih baik dari performa model Support Vector Machine yang dijadikan sebagai rujukan pada website baru yang belum pernah dilihat sebelumnya.

Model selanjutnya adalah model klasifikasi binari tanpa class weight yang dilatih hanya menggunakan dataset website yang sama saja. Sehingga terdapat lima total model yang dilatih khusus untuk masing-masing Detik.com, Iwanbanaran.com, Kompas.com, Teknojurnal.com, dan Tipspintar.com. Performa variasi model per website ini dapat dilihat pada Tabel IX.

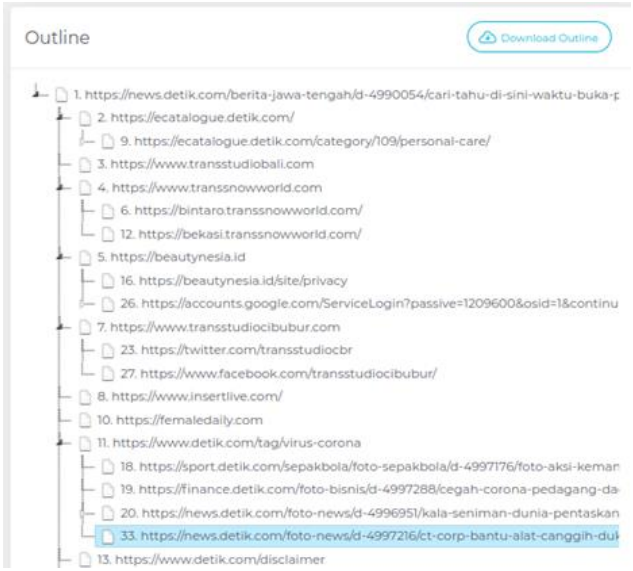
TABEL IX
PERFORMA MODEL KLASIFIKASI BINARI PER WEBSITE

Website	**Precision**	**Recall**	**F-1 Score**
Detik.com	98.57%	98.57%	98.57%
Iwanbanaran.com	98.61%	100%	99.3%
Kompas.com	95.76%	91.87%	93.78%
Teknojurnal.com	99.03%	100%	99.51%
Tipspintar.com	97.82%	98.15%	97.99%

Model neural network klasifikasi binari pada tiap website tidak dapat menghasilkan performa sempurna seperti model klasifikasi SVM. Hal ini mungkin dikarenakan tingkat kesensitifan model neural network terhadap noise sehingga mencegah performa model untuk mencapai nilai sempurna. Selain itu, dalam konteks satu model neural network per website, performa model neural network klasifikasi binari tidak jauh beda dengan performa model neural network klasifikasi multiclass. Oleh karena itu, jika terdapat suatu kebutuhan untuk melatih sebuah model neural network untuk suatu website, maka model neural network klasifikasi multiclass adalah pilihan yang lebih baik.

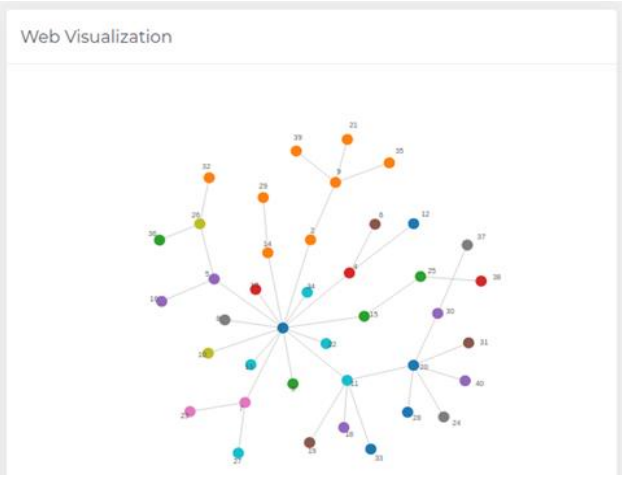
G. Web Data Visualization

Visualisasi outline tree dari hasil web crawling berhasil ditampilkan pada bagian kiri halaman hasil utama. Halaman web yang didapatkan dari halaman web lainnya akan berada di dalam turunan dari halaman web tersebut dengan posisi sedikit menjorok ke kanan. Turunan suatu halaman web juga dapat di buka tutup dengan klik. Hasil visualisasi grafik outline tree treeJS dapat dilihat pada Gambar 5 di bawah ini.



Gambar. 5. Outline Tree

Visualisasi grafik D3 dari hasil web crawling berhasil ditampilkan pada bagian tengah halaman hasil utama. Setiap node memiliki banyak koneksi ke halaman web yang lain. Setiap node juga memiliki warna yang berbeda-beda tergantung pada domain dari node tersebut. Hasil visualisasi grafik D3 dapat dilihat pada Gambar 6 di bawah ini.



Gambar. 6. Grafik

Blok teks terklasifikasi dari hasil klasifikasi neural network classifier ditampilkan pada bagian kanan halaman hasil utama. Blok teks yang diklasifikasikan sebagai judul akan ditampilkan sebagai judul dengan tag heading 1. Blok teks yang diklasifikasikan sebagai penulis akan ditampilkan sebagai penulis dengan tag paragraph dan emphasize. Blok teks yang diklasifikasikan sebagai konten akan ditampilkan sebagai konten dengan tag paragraph. Tampilan blok terklasifikasi dapat dilihat pada Gambar 7 di bawah ini.

Gambar. 7. Blok Teks Terklasifikasi

VI. KESIMPULAN

Website Web Content Extractor dapat mengotomatisasi proses pengambilan artikel di internet dengan menggunakan kombinasi web crawler dan ekstraktor blok teks dan CSS. Namun untuk mendapatkan hasil penyaringan konten murni yang ideal, model neural network yang digunakan harus dilatih dengan website yang akan diambil artikelnya. Dengan sampel yang cukup, model neural network dapat mengidentifikasi judul, penulis, dan konten artikel.

Website Web Content Extractor dapat memvisualisasi berbagai halaman website dengan mempresentasikan setiap halaman web sebagai sebuah node yang memiliki banyak koneksi ke halaman web yang lain. Website ini juga memiliki visualisasi berupa outline tree.

Model Deep Feed Forward Neural Network memiliki potensi untuk melakukan klasifikasi antara konten dan non konten secara umum selama diberi dataset latihan yang cukup dan bervariasi. Namun model ini belum dapat memberikan performa yang cukup untuk mengklasifikasikan judul dan penulis pada website secara umum.

Model Deep Feed Forward Neural Network dapat melakukan klasifikasi antara konten dan non konten untuk halaman web secara umum dengan lebih baik dibandingkan menggunakan Support Vector Machine. Performa F-1 score yang dihasilkan model neural network dua kali lebih baik ketimbang performa F-1 score model SVM. F-1 score dari model neural network adalah 62.87%. Sedangkan F-1 score dari model Support Vector Machine adalah 31.28%.

Penggunaan DBSCAN tidak cocok untuk mengklasterisasi blok teks pada halaman website yang memiliki berbagai template dan tidak menggunakan prinsip atau atribut yang sama untuk setiap elemen artikelnya. Penggunaan algoritma LCS hanya cocok digunakan ketika semua teks blok yang berisi konten berhasil diklasterisasi dalam satu klaster saja. Selain itu, meta tag yang deskriptif juga sangat diperlukan pada setiap halaman berartikel agar proses penilaian klaster berisi teks artikel dapat bekerja dengan baik. Website Indonesia yang menjadi data set sayangnya tidak menggunakan meta tag dengan standard yang disebutkan diatas.

DAFTAR PUSTAKA

[1] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[2] F. Chollet, "Keras: The Python Deep Learning library," *Keras.Io*, 2015.
[3] M. Abadi *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv Prepr. arXiv1603.04467*, 2016.
[4] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 data-driven documents," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2301–2309, 2011.
[5] Z. Zhou and M. Mashuq, "Web content extraction through machine learning," *Standford Univ.*, pp. 1–5, 2014.
[6] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[7] E. Lim, E. I. Setiawan, and J. Santoso, "Stance Classification Post Kesehatan di Media Sosial Dengan FastText Embedding dan Deep Learning," *J. Intell. Syst. Comput.*, vol. 1, no. 2, pp. 65–73, 2019.
[8] M. A. Rahman, H. Budianto, and E. I. Setiawan, "Aspect Based Sentimen Analysis Opini Publik Pada Instagram dengan Convolutional Neural Network," *J. Intell. Syst. Comput.*, vol. 1, no. 2, pp. 50–57, 2019.
[9] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, and others, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *kdd*, 1996, vol. 96, no. 34, pp. 226–231.
[10] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," 2015.

Sentiment Classification untuk Opini Berita Sepak Bola

Eka Rahayu Setyaningsih, Program Studi Informatika Institut Sains dan Teknologi Terpadu Surabaya

***Abstrak*— Pada penelitian ini akan dibahas mengenai sebuah aplikasi yang dibuat secara khusus untuk mengkategorikan opini masyarakat terhadap sebuah berita Sepak Bola. Opini yang diolah diperoleh dari dua sumber, yaitu melalui hasil crawl situs berita olah raga dan opini yang ditambahkan oleh user sendiri pada aplikasi ini. Opini yang ada nantinya akan disajikan secara terpisah menurut kelompoknya; sentiment positive, negative, maupun netral. Proses klasifikasinya sendiri terdiri dari dua tahap. Tahap pertama adalah proses preprocessing yang terdiri atas proses tokenisasi, normalisasi, case folding, stop word removing, common word removing, stemming. Tahap kedua adalah mengklasifikasikan opini-opini tersebut dengan algoritma Baseline, dan Naive Bayes. Opini yang digunakan untuk proses klasifikasi yaitu opini yang menggunakan bahasa Inggris dari situs fifa.com dan goal.com. Dari perhitungan macroaveraged untuk setiap kelas, didapatkan akurasi 93,06%, presisi 81,90%, dan recall 92,67% untuk kelas sentiment positive. Dari perhitungan kelas sentiment negative didapatkan akurasi 87,73%, presisi 96,29%, dan recall 83,63%. Dari perhitungan kelas sentiment netral didapatkan akurasi 92,26%, presisi 64,44%, dan recall 90,37%. Kesimpulan yang diperoleh saat penelitian ini dari awal hingga akhir adalah, proses crawling yang digunakan untuk mendapatkan berita dan komentar berita sangat membantu dalam penambahan konten website, tetapi banyak sekali komentar berita yang diperoleh kurang cocok untuk proses klasifikasi.**

***Kata Kunci*—Sentiment Analysis, Opinion Classification, Naive Bayes, Football.**

I. PENDAHULUAN

Informasi yang tersedia dalam dunia web dibagi menjadi dua tipe yaitu fakta dan opini. Setiap orang bebas untuk mengekspresikan opininya dalam berbagai ragam topik dan melalui berbagai macam media online seperti blog, jejaring sosial dan lain-lain. Salah satu bentuk penyajian informasi tekstual yaitu Sentiment Analysis.

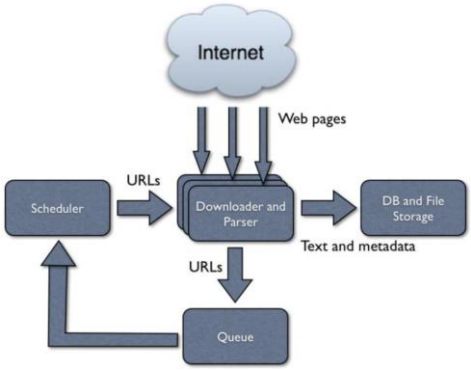
Belakangan, sentiment analysis menarik sebagian besar perhatian baik dari akademis maupun industri [1]–[5]. Hal ini disebabkan karena banyaknya masalah penelitian dan berbagai macam aplikasi yang menantang. Dengan menggunakan sentiment analysis, seseorang dapat melihat tanggapan seseorang terhadap sesuatu permasalahan yang diamati sebelum menyimpulkan sebuah keputusan.

Berita sepak bola adalah salah satu topik yang sering digunakan oleh seseorang untuk memberikan opini. Opini yang diberikan bisa sebagai opini positif atau opini negatif maupun opini netral. Untuk dapat mengetahui opini tersebut sebagai opini positif atau negatif atau netral, pengguna harus membaca dan mengamati opini tersebut satu per satu. Oleh karena itu, untuk mempermudah pengguna mengetahui opini tersebut masuk ke opini positif atau negatif atau netral.

II. TINJAUAN PUSTAKA

A. *Web Crawling*

Web Crawling merupakan suatu proses yang digunakan untuk menjelajah dan mengambil sekumpulan halaman dari sebuah web. Untuk melakukan proses tersebut dibutuhkan sebuah perangkat lunak yang disebut web crawler. Web crawler adalah salah satu komponen penting dalam mesin pencari modern seperti Google, Yahoo, dan lain-lain. Web crawler sering dikenal dengan nama web spider atau web robot.



Gambar 1. Cara Kerja Crawler

Seperti yang ditunjukkan pada gambar 1, mula-mula web crawler akan memulai kerjanya dengan mengunjungi situs yang sudah disebutkan oleh user sebagai *url seed* atau URL yang ada di dalam sebuah database yang digunakan untuk menyimpan alamat situs yang ingin dikunjungi. Database penyimpanan alamat situs tersebut dikenal dengan sebutan frontier url.

Setelah web crawler tiba atau sampai pada halaman website yang ditentukan, maka web crawler akan melakukan proses fetching. Proses fetching adalah proses yang dilakukan oleh web crawler untuk mengambil dokumen HTML yang terdapat dalam suatu halaman website. Hasil fetching web crawler yang berupa data dan meta data akan disimpan ke dalam penyimpanan utama. Sedangkan hasil penguraian web crawler berupa outlink atau tautan lain yang didapat, akan dimasukkan ke dalam queue.

Bersamaan dengan berjalannya proses crawling, seluruh

Eka Rahayu Setyaningsih, Departemen Informatika, Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Indonesia (e-mail: eka@stts.edu)

outlink yang didapat dari sebuah halaman website akan didaftarkan dalam sebuah queue yang berupa daftar outlink yang ada. Setelah proses fetching pada halaman tersebut selesai, web crawler akan melakukan proses fetching kembali. Selanjutnya, proses fetching akan dilakukan pada halaman web lain yang telah terdaftar dalam queue tersebut. Apabila dalam proses fetching selanjutnya terdapat outlink yang telah terdapat dalam queue, maka outlink tersebut tidak akan diproses dan ditambahkan ke dalam queue. Proses ini akan dilakukan web crawler sampai outlink yang terdaftar dalam queue habis atau batasan kedalaman crawling yang ditentukan oleh user.

B. Sentiment Classification

Sentiment classification merupakan suatu proses mengklasifikasian teks yang terdapat pada suatu bacaan atau dokumen. Sentiment classification adalah sebuah teknik pengklasifikasian yang mana pembelajarannya menggunakan metode supervised learning. Target output dari metode ini biasanya diambil dari suatu dataset yang telah ada.

Sentiment classification bertujuan untuk menentukan tanggapan dan sikap seseorang sehubungan dengan beberapa topik [6], [7]. Sering kali, sentiment classification digunakan untuk menentukan apakah penulis menyukai atau tidak menyukai ebuah produk dari review yang mereka tulis. Input dari sentiment classifier selalu sebuah teks beropini. Teks beropini adalah teks yang mengandung sentimen positif, atau negatif. Sedangkan output dari sentiment classifier yaitu penggolongan teks ke sebuah kelas seperti positif atau negatif.

Untuk menyelesaikan permasalahan dari Sentiment Analysis, ada banyak algoritma yang dapat digunakan seperti algoritma Naive Bayes, Support Vector Machine, C45, K-Nearest Neighbor, K-Means, Maximum Entropy dan lain-lain. Algoritma yang paling sering digunakan yaitu Naive Bayes dan Support Vector Machine [8]. Naive Bayes sering digunakan karena sederhana tetapi memiliki akurasi yang tinggi sedangkan Support Vector Machine dikarenakan pengerjaan yang sangat baik pada data dengan banyak dimensi. Dalam penelitian ini algoritma yang digunakan yaitu algoritma Naive Bayes dan algoritma Baseline.

C. Baseline Algorithm

Algoritma Baseline [9] melakukan analisis sentimen pada komentar-komentar dengan menggunakan daftar kata positif dan negatif. Algoritma ini akan membandingkan setiap kalimat yang terdapat satu komentar. Kata-kata yang terdapat pada setiap kalimat hanya dibandingkan dengan daftar yang ada dan dikelompokan langsung tanpa diproses terlebih dahulu kebenarannya kalimat tersebut positif atau negatif.

Segmen Program 1- Algoritma Baseline

```
1.  L1={List of Positif words}
2.  L2={List of Negatif words}
3.  for each commentc in database
    do
4.    for each words w in c do
5.       increment count of positif words
         in L1 that contained in w
6.       increment count of negatif words in L2
         that contained in w
7.    end
8.    if count of negatif word > count of positif
      word
9.       return "negatif"
10.   else return "positif"
11. end
```

Metode Baseline yang akan digunakan dalam penelitian ini adalah supervised baseline. Supervised baseline adalah metode baseline yang mana daftar kata-kata positif dan negatif yang digunakan untuk perbandingan sudah tersedia dan tidak diperlukannya pembelajaran secara otomatis. Daftar kata positif yang digunakan seperti adorable, capable, cool, fair, dan lain-lain. Daftar kata negatif yang digunakan seperti arrogant, awful, selfish, rough, careless, dan lain-lain.

Berikut adalah contoh komentar yang akan dibandingkan dengan daftar kata-kata positif dan negatif yang ada.

Contoh komentar:
"I agree with him on most things, but Ferguson was a great coach, The coach has to have a bigger personality than the players at a big club and he definitely was, That's why they were so successful. Moving on to england now, he's absolutely right, The media and pundits overhype every little talent".

Hasil analisa komentar:
- Kata Positif : 5 buah
- Kata Negatif : 1 buah

Output: Komentar 1 termasuk komentar positif.

TABEL I
DAFTAR KATA POSITIF DAN NEGATIF

Daftar Kata Positif	Daftar Kata Negatif
Abound	abnorm
Absolut	abolish
Absorb	abomin
Abstemiou	abort
Abund	abrad
Abundance	abras
Accept	abrupt

Dengan terlebih dahulu membentuk daftar kata negative dan daftar kata positif yang akan digunakan selama penelitian (seperti yang ditunjukkan pada table 1), maka dapat disimpulkan bahwa dari komentar pertama dapat diketahui bahwa komentar tersebut termasuk komentar positif karena jumlah kata positif lebih banyak dari jumlah kata negatif. Kata-kata yang termasuk kata positif yaitu agree, great, bigger, successfull, dan right. Kata-kata yang termasuk kata negatif yaitu overhype.

D. Naive Bayes Algorithm

Naive Bayes Classifier (NBC) merupakan metode pembelajaran dengan konsep probabilitas sederhana. NBC menggunakan teorema kuno, warisan abad ke-18, yang ditemukan oleh Thomas Bayes. NBC menyertakan dokumen klasifikasi terbimbing, metode pembelajaran yang menghasilkan fungsi untuk memetakan masukan ke keluaran yang diinginkan. NBC menganggap kemunculan satu kata tidak mempengaruhi kemunculan kata lainnya. NBC mampu memberikan kinerja yang cukup baik untuk banyak kasus

modern dengan data yang besar. Adapun untuk menghitung probabilitas fitur kata menggunakan persamaan (1):

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|} \quad \text{..... (1)}$$

Kemudian untuk menghitung probabilitas prior menggunakan persamaan (2):

$$\hat{P}(c) = \frac{N_c}{N} \quad \text{..... (2)}$$

Terakhir untuk menentukan sentimen menggunakan persamaan (3):

$$c = \text{argmax}_c \hat{P}(c) \prod_i \hat{P}(w_i \mid c) \quad \text{..... (3)}$$

E. Language Identification

Language identification merupakan suatu proses identifikasi bahasa. Proses identifikasi bahasa ini bertujuan untuk mengetahui sebuah teks yang tertulis termasuk ke dalam golongan bahasa tertentu. Teks yang teridentifikasi bisa termasuk bahasa Inggris, Indonesia atau bahasa-bahasa lainnya. Proses pengidentifkasian bahasa ini, akan mencocokan setiap kata yang terdapat pada suatu teks dengan database yang ada. Berikut adalah contoh pengidentifikasian sebuah teks. Contoh teks yang akan diidentifikasi yaitu "worst player ever". Setelah diidentifikasi, teks ini digolongkan dalam bahasa Inggris. Setiap pengidentifikasian akan memiliki nilai angka yang menunjukkan ketepatan pengindentifikasian bahasa. Jika suatu teks digolongkan ke dalam 2 bahasa, maka nilai pengidentifikasian bahasa yang tertinggi akan menjadi penentu bahasa.

Penggunaan Language Detection API dapat dilakukan dengan 2 cara. Cara pertama yaitu menggunakan API yang telah disediakan untuk beberapa bahasa pemrograman seperti Ruby, Java, Python, dan PHP. Karena bahasa pemrograman yang digunakan dalam penelitian ini tidak terdaftar, maka akan menggunakan cara yang lain. Cara yang lain yaitu dengan metode GET dan POST request. Metode POST lebih disarankan untuk jumlah request yang lebih banyak. Kedua metode ini akan memberikan hasil dalam format json. Cara penggunaan GET dan POST request ini cukup mudah, dengan mengakses ke alamat url http://ws.detectlanguage.com/0.2/detect?parameter. Ada dua parameter yang digunakan yaitu q dan key. Berikut penjelasan parameter yang digunakan:

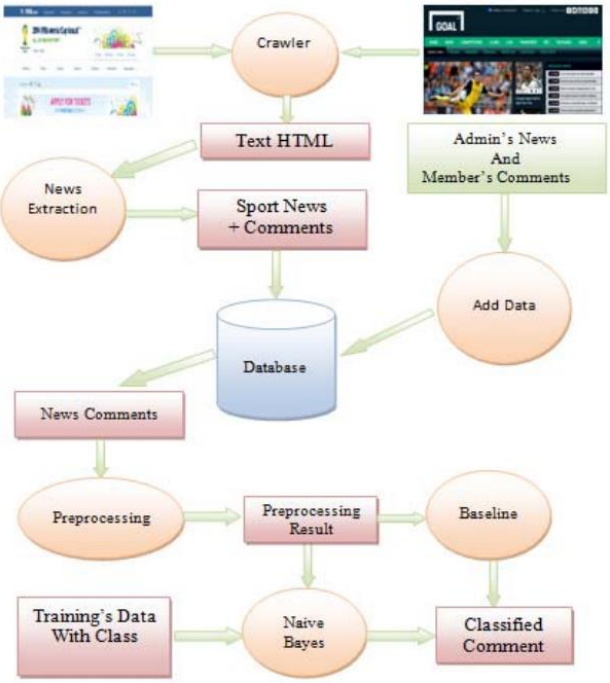
1. Parameter q digunakan untuk memasukan sebuah kata atau text yang akan di deteksi. Tidak ada batasan karakter untuk parameter ini.
2. Parameter key ini akan diisi dengan API key yang didapat setelah melakukan pendaftaran pada detectlanguage.com

Berdasarkan kedua parameter tersebut untuk menggunakan GET request dapat dilihat pada URL berikut: "http://ws.detectlanguage.com/0.2/detect?q=buen os+dias+senor&key=demo".Dari contoh tersebut, berarti teks yang diingin dideteksi adalah buenos dias senor dan key yang digunakan adalah demo. Jika URL itu dijalankan pada web browser maka akan menampilkan sebuah halaman yang berisi hasil pendeteksian bahasa. Setiap hasil pendeteksian akan berisi kode dari bahasa yang digunakan, nilai keakuratan bahasa, dan status bahasa yang terdeteksi dapat dipercaya atau tidak. Language Detection API untuk sementara mendeteksi sebanyak 83 bahasa.

Untuk mendeteksi beberapa teks, dapat dilakukan dalam satu URL. Berikut adalah contoh pendeteksian: http://ws.detectlanguage.com/0.2/detect?q []=buenos+dias &q[]=morning&key=demo. Dalam pendeteksian tersebut, setiap parameter q yang digunakan akan dihitung satu parameter. Jadi untuk pendeteksian tersebut, terdapat dua request yang dilakukan. Hasil request yang dilakukan akan sesuai dengan urutan parameter q yang dikirimkan. Dengan melakukan banyak request dalam satu URL menghemat kuota jaringan dikarenakan jumlah kuota yang terbatas. Untuk penelitian ini digunakan Free Plans untuk language detection API yang terbatas untuk 5000 request dan 1MB kuota perharinya.

III. ARSITEKTUR SISTEM

Gambaran arsitektur sistem secara singkat namun menyeluruh dan menggambarkan keseluruhan fase yang ada, tahapan serta proses yang dilakukan dapat dilihat pada gambar 2.



Gambar 2. Arsitektur Sistem

Arsitektur sistem yang dibangun di penelitian ini dibagi menjadi 3 fase yaitu fase ekstraksi data, pembentukan dataset dan fase klasifikasi data dengan hasil belajar dari dataset yang telah dibentuk pada fase sebelumnya.

A. Fase Ekstraksi Data

Pada fase ekstraksi data, sistem akan mencari informasi berita sepak bola tersebut melalui crawling situs goal.com/en dan fifa.com. Dari setiap halaman yang berhasil dikunjungi oleh crawler, akan didapatkan halaman-halaman berita. Setiap halaman berita yang didapatkan pada tahap ini masih berupa sebuah dokumen HTML yang memuat banyak tag-tag

HTML yang perlu dibersihkan untuk mendapatkan informasi penting dari berita tersebut seperti judul, isi dan informasi-informasi lain terkait berita.

```
<article>
<a    href="/en/news/12/spain/2013/10/02/4305154/real-madrid-handed-
bale-blow">
<img    src="http://u.goal.com/322600/322604_sq_thumb.jpg"    alt=""
width="72" height="72">
</a>
<h3><a href="/en/news/12/spain">Spain</a></h3>
<h2><a       href="/en/news/12/spain/2013/10/02/4305154/real-madrid-
handed-bale-blow">Bale out of Copenhagen clash</a></h2>
<!-- module:SocialShareIcons -->
<div class="module module-social-share-icons dark"
data-role="share-icons"
data-
url="http://www.goal.com/en/news/12/spain/2013/10/02/4305154/real-
madrid-handed-bale-blow"
data-title="Bale out of Copenhagen clash"
data-description="The   24-year-old   has   suffered   a   stop-start
beginning to his Blancos career and will miss the Liga side's
latest European tie"
data-image="http://u.goal.com/322600/322604_hp.jpg"
```

Gambar 3. Contoh Hasil Proses Fetching

Selanjutnya pada dokumen HTML yang diperoleh dilakukan ekstraksi informasi id, judul, sub judul, tanggal, tag, isi berita, sumber, hingga link gambar terkaitnya dengan menggunakan regex, hingga diperoleh informasi seperti yang ditunjukkan pada gambar 4 berikut ini.

```
Kode Berita       :  FF2261338
Judul Berita      :  Matri  moved  to  Fiorentina  to  fix
           injury crisis
Sub Judul Berita  : Alessandro Matri has been loaned out
           to Fiorentina from rivals AC Milan on a six-
           month contract to help fill La Viola's void up
           top with strikers Gius...
Isi Berita        : Italian Serie A side Fiorentina moved
           to  resolve  an  injury  crisis  affecting  their
           strikers  by  signing  Italian  international
           striker Alessandro Matri on loan from rivals
           left  by  injuries  to  Giuseppe  Rossi,  leading
           scorer  in  Serie  A  with  14  goals,  and  German
           international  Mario  Gomez."There  are  some
           minor details to eke out and then I will be
           available for the coach,"
Pembuat Berita    : Fifa
Sumber            : Fifa
Tag Berita        : Italy
URL foto Berita   :
           www.fifa.com/mm/photo/tournament/competition/0
           1/69/73/08/1697308_small-lnd.jpg
Tanggal Berita    : 2013-10-10
```

Gambar 4. Contoh Hasil Ekstraksi Informasi

Proses selanjutnya yang dilakukan adalah melakukan ekstraksi setiap teks komentar yang diberikat terhadap artikel berita terkait. Proses yang dilakukan sama seperti proses ekstrasi informasi berita, yaitu dengan menggunakan regex. Setelah berhasil mendapatkan informasi tersebut, sistem akan melakukan filter untuk mendapatkan komentar yang menggunakan bahasa Inggris saja. Untuk membedakan komentar tersebut menggunakan bahasa Inggris atau tidak, cara yang akan digunakan yaitu dengan menggunakan language detector API.

B. *Fase Pembentukan Dataset*

Karena Naïve Bayes merupakan algoritma yang termasuk dalam kategori supervised learning, maka terlebih dahulu dilakukan pembentukkan dataset. Dataset ini penting digunakan untuk proses training algoritma Naïve Bayes itu sendiri. Penggunaan algoritma Naïve Bayes juga mengharuskan statistika persebaran dataset untuk setiap kata berada pada posisi yang seimbang.

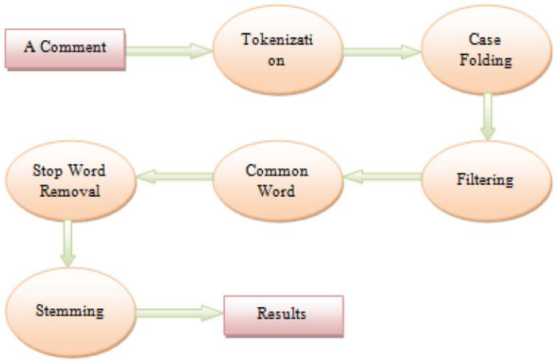
Dataset yang digunakan pada penelitian ini diambil dari hasil ekstraksi komentar yang didapat dari situs yang sudah ditentukan yaitu fifa.com dan goal.com/en. Kemudian dataset tersebut diberi label secara manual yaitu positif, negatif atau netral. Positif dilambangkan dengan P, Negatif dilambangkan dengan N, sedangkan Netral dilambangkan dengan O. Dataset yang digunakan berjumlah 20.000 komentar. Selanjutnya dataset ini akan diproses melalui preprocessing. Pada program sesungguhnya dataset yang digunakan berjumlah 500 komentar. Contoh bentuk dataset yang digunakan sebagai data training dapat dilihat pada gambar 5 berikut ini.

```
O     Well said.
P     Glad to see an Englishman speaking the truth.
N     wanker would be appropriate
P     As much as I hate this guy, I gotta agree with him for
      once!
P     lmao i like this guy
N     Barton is a twat but he's spot on here about the national
      team lol. However his criticism of SAF is uncalled for.
P     lol love this guy says what most people think
N     England is a retarted footballing nation at the core.
      seriously..... no English manager has ever won the PL....
      no other league is that depended on foreigners
O     When was england ever good ?
O     When they could cheat
O     Ferguson is not from England....
N     Barton is so raw, he is just straight up badass
O     true, how many British Players are being Exported to
      other countries? definitely not as much as the Spanish or
      Brasilian or even mexican players.
N     And you would know crap wouldn't you Barton.
N     Barton is an ass, that being said he kinda has point even
      if he just is an ass
P     wow. Joey is such a fearless lad. finally he made great
      point and I totally agree with him. England need a change
      seriously.
P     Good bang
```

Gambar 5. Contoh Dataset

C. *Fase Klasifikasi Data*

Proses Preprocessing dimulai dengan melakukan proses tokenisasi pada sebuah komentar. Proses tokenisasi adalah proses untuk memecah suatu komentar menjadi sekumpulan kata. Biasanya proses ini akau memecah suatu dokumen berdasarkan karakter spasi, atau tanda koma.



Gambar 6. Preprocessing

Proses selanjutnya case folding yaitu proses yang digunakan untuk penyamaan case huruf. Filtering adalah proses untuk menghilangkan karakter yang tidak digunakan dalam penentuan sebuah kelas sentimen. Karakterkarakter yang akan dihilangkan seperti kata yang mengandung perulangan huruf lebih dari tiga, tanda baca, symbol.

Common word dan stop word removal adalah proses untuk menghilangkan kata-kata yang tidak dipakai dalam proses penentuan sebuah kelas sentimen. Untuk common word removal, kata-kata yang dihilangkan adalah kata-kata yang ada hubungan dengan topik penelitian, seperti nama pemain bola, nama stadiun, nama pelatih, nama klub, nama negara, singkatan dan lain-lain. Sedangkan untuk proses stop word removal, katakata yang dihilangkan antara lain kata hubung, kata ganti orang, kata benda, dan lain-lain.

Stemming adalah proses yang digunakan untuk mengembalikan sebuah kata menjadi kata dasarnya. Metode stemming yang akan digunakan yaitu porter stemmer [10]. Metode ini ditemukan oleh Martin F. Porter. Metode ini paling sering digunakan dalam proses stemming untuk bahasa Inggris.

Setelah proses preprocessing selesai, proses training data baru dilakukan. Berikut adalah contoh hasil proses training data.

TABEL 2
DAFTAR STATISTIK DAN PROBABILITAS

Words	Jumlah			Probabilitas		
	Pos	Neg	Net	Pos	Neg	Net
well	3	3	0	0.00300	0.00207	0.00088
said	3	4	1	0.00300	0.00259	0.00176
glad	1	0	0	0.00150	0.00051	0.00088
hate	1	3	0	0.00150	0.00207	0.00088
gotta	1	0	0	0.00150	0.00051	0.00088
agree	3	0	0	0.00300	0.00051	0.00088
like	4	7	3	0.00375	0.00414	0.00353
love	3	2	0	0.00300	0.00155	0.00088
retart	0	1	0	0.00075	0.00103	0.00088

Setelah melakukan perhitungan proses training dan perhitungan probabilitas dari training data tersebut dapat dilakukan proses pengklasifikasian sentimen. Berikut adalah contoh proses pengklasifikasian sentimen.

- Kata yang ingin diklasifikasi:
  i don't like barton and england. Both of them are retarted.
- Kata setelah preprocessing: like dan retart
- Perhitungan Probabilitas Positif:
  0.00375 * 0.00075 = 0.0000028125
- Perhitungan Probabilitas Negatif:
  0.00414 * 0.00103 = 0.0000042642
- Perhitungan Probabilitas Netral:
  0.00353 * 0.00088 = 0.0000031064

Dari perhitungan ketiga probabilitas tersebut, kata yang ingin diklasifikasi dapat digolongkan ke sentimen negatif karena nilai probabilitas negatif lebih besar dari nilai probabilitas positif dan probabilitas netral.

IV. INPUT OUTPUT SISTEM

Pada tahap ini akan dijelaskan mengenai input dan output dari sistem ini. Sistem akan diberi inputan yang kemudian akan diproses untuk mendapatkan output. Berikut akan dibahas input dan output sistem beserta contoh-contoh yang relevan.

A. *Input*

Input yang diperlukan sistem adalah kalimat dalam bahasa Inggris yang baik dan benar tentang komentar berita sepak bola. Kalimat yang diberikan tidak boleh memiliki singkatan, perulangan huruf yang tidak sesuai atau kata yang tidak baku. Contoh kalimat input yang baik:

- yes you are the best but only for barcelona because you not going anywhere but christiano ronaldo proof he is the best alltime anywhere any place
- I love messi ! No one can beat him!

B. *Output*

Sistem memberikan beberapa jenis output sehubungan dengan input teks yang diberikan oleh pengguna. Pertama, output berupa penggolongan bahasa yang didapat menggunakan Language Detection API. Ada kemungkinan sistem mendapatkan satu atau lebih jenis penggolongan bahasa. Output yang kedua adalah pengklasifikasian input teks berdasarkan kedua algoritma yang digunakan. Untuk algoritma naive bayes, input teks akan digolongkan ke sentimen positif, negatif, atau netral. Untuk algoritma Baseline, input teks hanya digolongkan ke sentimen positif atau negatif

V. UJICOBA

Uji coba sistem meliputi uji coba akurasi output yang dilakukan dengan menghitung hasil akurasi, precision dan recall [11] hasil identifikasi 150 komentar berita yang didapatkan baik melalui crawling atau penambahan komentar oleh member. Dari perhitungan macroaveraged untuk setiap kelas, didapatkan akurasi 93,06%, presisi 81,90%, dan recall 92,67% untuk kelas positif. Dari perhitungan kelas negatif didapatkan akurasi 87,73%, presisi 96,29%, dan recall 83,63%. Dari perhitungan kelas netral didapatkan akurasi 92,26%, presisi 64,44%, dan recall 90,37%.

VI. KESIMPULAN

Kesimpulan yang diperoleh saat penelitian ini dari awal hingga akhir adalah.

1. Proses crawling yang digunakan untuk mendapatkan berita dan komentar berita sangat membantu dalam penambahan konten website, tetapi banyak sekali komentar berita yang diperoleh kurang cocok untuk proses klasifikasi. Oleh karena itu dibutuhkan penyaringan komentar sehingga didapatkan komentar yang lebih baik.
2. Agar hasil proses preprocessing menjadi lebih baik, sistem membutuhkan daftar stopword dan commonword yang lebih banyak.Semakin baik hasil dariproses preprocessing akan membantu proses klasifikasi data.
3. Untuk meningkatkan kinerja dari proses klasifikasi sentiment diperlukan training data yang lebih banyak. Dengan hasil training yang lebih banyak diharapkan akan mendapatkan akurasi yang lebih baik.
4. Penggunaan Language Detection API untuk pendeteksian bahasa sangat membantu dalam penentuan komentar yang digunakan untuk proses klasifikasi, tetapi sebagian pendeteksian yang dilakukan tidak menghasilkan bahasa yang sesuai.

DAFTAR PUSTAKA

[1] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 168–177, doi: 10.1145/1014052.1014073.

[2] B. Liu and others, "Sentiment analysis and subjectivity.," *Handb. Nat. Lang. Process.*, vol. 2, no. 2010, pp. 627–666, 2010.

[3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," *arXiv Prepr. cs/0205070*, 2002.

[4] B. Pang, L. Lee, and others, "Opinion mining and sentiment analysis," *Found. Trends® in Inf. Retr.*, vol. 2, no. 1--2, pp. 1–135, 2008.

[5] M. A. Rahman, H. Budianto, and E. I. Setiawan, "Aspect Based Sentimen Analysis Opini Publik Pada Instagram dengan Convolutional Neural Network," *J. Intell. Syst. Comput.*, vol. 1,

no. 2, pp. 50–57, 2019.
[6] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, “Predicting the political alignment of twitter users,” in *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, 2011, pp. 192–199.
[7] A. Makazhanov, D. Rafiei, and M. Waqar, “Predicting political preference of Twitter users,” *Soc. Netw. Anal. Min.*, vol. 4, no. 1, pp. 1–15, 2014.
[8] N. W. S. Saraswati, “Text mining dengan metode na{\"\i}ve bayes classifier dan support vector machines untuk sentiment analysis,” *Univ. UDAYANA, Tek. Elektro. Denpasar Univ. UDAYANA*, 2011.
[9] S. I. Wang and C. D. Manning, “Baselines and bigrams: Simple, good sentiment and topic classification,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2012, pp. 90–94.
[10] M. F. Porter, “An algorithm for suffix stripping,” *Program*, 1980.
[11] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*, vol. 39. Cambridge University Press Cambridge, 2008.

Tamagotchi Augmented Reality yang Dilengkapi dengan Mini Games

Hendrawan Armanto, *Informatika Institut Sains dan Teknologi Terpadu Surabaya*,
Edwin Sidharta, *Informatika Institut Sains dan Teknologi Terpadu Surabaya*

***Abstrak*— Pada saat ini, teknologi mobile telah berkembang dengan pesat. Dalam kesehariannya, manusia tidak dapat lepas dari handphone. Hal ini menyebabkan munculnya berbagai aplikasi dan game yang bertujuan tentu saja untuk membantu ataupun memberikan kesenangan kepada penggunanya. Saat ini perkembangan game, juga sangat pesat dan telah mencapai titik dimana berbagai jenis game dikembangkan. Tidak hanya berhenti pada perkembangan jenis game, bahkan cara bermain dari game itu sendiri juga ikut berkembang. Yang dulunya permainan mobile dilakukan secara virtual, saat ini permainan sudah menyentuh area Augmented Reality (AR) dimana pemain dapat melihat benda-benda tidak nyata (buatan) dalam dunia nyata (dunia manusia). Walaupun permainan AR semakin berkembang, tetapi masih sedikit permainan AR bergenre Virtual Pet. Penelitian ini bertujuan untuk mengembangkan Permainan Virtual Pet dan mengukur tingkat kesenangan dalam memainkan permainan ini. Permainan dikembangkan dengan menggunakan Unity Game Engine dengan bantuan package AR Foundation dan penyimpanan data pada Firebase. Ujicoba akan dilakukan kepada 40 orang (pria dan wanita) pemain game yang pernah bermain virtual pet sebelumnya. Hasil akhir ujicoba menunjukan bahwa dalam segi teknis permainan berjalan dengan baik dan disukai oleh pemain akan tetapi ada sebagian pemain yang tingkat kesenangannya rendah cenderung menengah hal ini dikarenakan gambar monster yang digunakan kurang menarik dan kurangnya fitur terkait monster itu sendiri.**

***Kata Kunci*— Game, Markerless Augmented Reality, Marker Based Augmented Reality, Virtual Pet.**

I. PENDAHULUAN

Perkembangan game saat ini telah sampai pada titik kemajuan yang pesat. Berbagai jenis permainan telah diciptakan. Salah satu teknologi yang digunakan untuk bermain game dan sedang berkembang pesat saat ini adalah mobile phone atau handphone. Game juga memiliki berbagai macam genre, salah satu genre dari game adalah virtual pet. Virtual pet merupakan salah satu genre game yang populer sekitar tahun 1995. Virtual pet merupakan genre game yang mengutamakan interaksi antara pemain dan peliharaan virtualnya.

Hendrawan Armanto, Departemen Informatika, Institut Sains dan Teknologi Terapan Surabaya, Surabaya, Jawa Timur, Indonesia (e-mail: hendrawan@stts.edu)

Edwin Sidharta, Departemen Informatika, Institut Sains dan Teknologi Terapan Surabaya, Surabaya, Jawa Timur, Indonesia (e-mail: edwin3@mhs.stts.edu)

Pada umumnya dalam game virtual pet juga terdapat berbagai macam mini-game, dimana setiap mini-game memiliki cara bermainnya masing-masing. Tidak seperti genre game yang lain, virtual pet pada umumnya cenderung memiliki fokus pada interaksi antar pengguna dan peliharaan virtualnya. Game yang akan dibuat akan menggunakan game engine unity dan bergenre virtual pet dengan target platform Android, Pengembangan game ini dilakukan untuk menarik minat pemain yang belum pernah mempunyai peliharaan atau yang ingin mencoba memiliki peliharaan virtual. Tujuan dari penelitian ini adalah untuk mengembangkan game berjenis virtual pet yang mengguna teknologi Augmented Reality pada platform Android dan mengukur tingkat kesenangan dari pemain game ini. Game ini akan memiliki genre virutal yang membuat pemainnya merasakan pengalaman memiliki peliharaan yang dapat berinteraksi di dunia nyata.

II. TINJAUAN PUSTAKA

Pada penelitian ini, digunakan beberapa komponen yang menunjang dalam pengembangan game virtual pet. Komponen yang digunakan untuk pembuatan aplikasi adalah: Unity Game Engine, Firebase, AR Foundation, Image Tracking, dan Plane Detection.

A. Unity Game Engine [1]

Unity (biasa yang kita kenal sebagai Unity3D) merupakan game engine dan IDE atau Integrated Development Environment untuk membuat media yang interaktif yang biasanya khusus untuk game. Game engine ini juga dapat mengolah gambar, grafik, suara, input, dan lain-lain.

Unity juga merupakan game engine yang bermultiplatform. Unity mampu di publish menjadi Standalone (Aplikasi berbasis dekstop), berbasis web, Android, IoS Iphone, XBOX, dan PS3. Walau bisa dipublikasi ke berbagai platform, Unity perlu lisensi untuk dapat dipublikasikan ke platform tertentu. Tetapi Unity menyediakan untuk free user dan bisa di publish dalam bentuk Standalone (.exe) dan web. Untuk saat ini Unity sedang di kembangkan berbasis AR (Augment Reality).

B. Firebase Realtime Database [2]

Firebase Realtime Database adalah database yang di-host di cloud. Data disimpan sebagai JSON (JavaScript Object Notation) dan disinkronkan secara realtime ke setiap klien yang terhubung. Ketika Anda membuat aplikasi lintasplatform dengan SDK Android, iOS, dan JavaScript,

semua klien akan menerima update data terbaru secara otomatis.

C. *AR Foundation [3]*

AR Foundation adalah sebuah package yang ada di Unity3D untuk memudahkan developer dalam membangun sebuah aplikasi augmented reality. AR Foundation membutuhkan Plugin ARCore XR dan merupakan SDK yang digunakan dalam pembuatan game ini. Sedangkan ARCore Extensions adalah package yang menyediakan fungsionalitas ARCore tambahan yang dapat digunakan dengan AR.

D. *Image Tracking*

Image tracking merupakan salah satu fitur yang terdapat pada AR Foundation yang berfungsi agar setiap gambar berisi marker yang dideteksi oleh kamera dapat dibuatkan GameObjects, sebelum dapat mendeteksi gambar yang berisi marker tersebut diperlukan suatu ImageLibrary yang digunakan untuk menyimpan semua gambar yang dapat dideteksi oleh fitur ini.

E. *Plane Detection*

Plane detection merupakan salah satu fitur yang terdapat pada AR Foundation yang berfungsi agar setiap plane yang terdeteksi dibuat dalam bentuk GameObjects. Terdapat tiga mode deteksi dalam plane detection yaitu horizontal, vertikal, atau keduanya.

F. *Penelitian Terdahulu*

Sebelum melakukan penelitian ini, kami mempelajari terlebih dahulu beberapa penelitian yang telah ada baik terkait augmented reality ataupun penelitian terkait virtual pet tersendiri. Penelitian terakit augmented reality penting kami pelajari untuk mengetahui seberapa jauh perkembangan augmented reality saat ini sehingga teknologi augmented reality yang kami gunakan tidak tertinggal. Berikut adalah beberapa penelitian terkait augmeted reality yang kami pelajari yaitu Agmented Reality berbasiskan Vuforia dan Unity untuk studi kasus Gedung M di Universitas Semarang [4], Augmented reality untuk memvisualisasikan objek 2D menjadi 3D[5], Augmented Reality untuk Salesman Canvasing [6], Augmented Reality Objek 3D [7], dan Pengenalan Jenis Laptop menggunakan Metode Makerless [8].

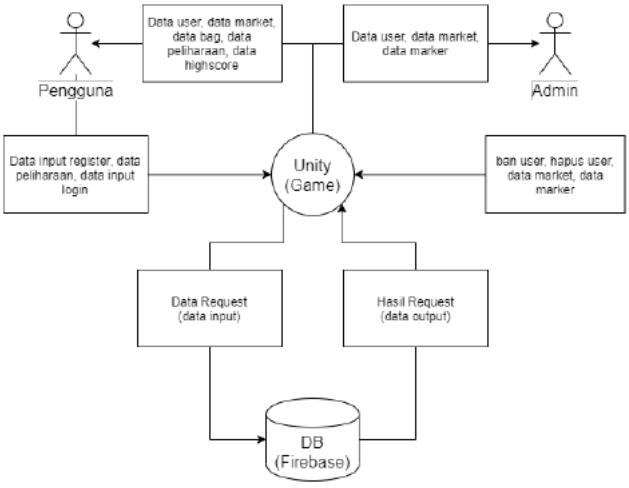
Sedangkan penelitian terkait virtual pet sendiri juga sama pentingnya untuk dipelajari sehingga penelitian di bidang ini terus berkembang dan tidak stagnan atau diulang terus menerus. Berikut beberapa penelitian terdahulu yang kami pelajari:

1. Virtual Pet Simulator Game using Augmented Reality on Android Plaform [9]
   Penelitian ini berfokus pada pengembangan virtual pet menggunakan augmented reality vuforia dimana virtual pet bertingkah laku seperti layaknya binatang peliharaan pada umumnya akan tetapi penelitian ini menyimpulkan bahwa tidak mungkin virtual pet menggantikan binatang peliharaan asli.
2. Eva: A Virtual Pet in Augmented Reality [10]
   Sama hal nya dengan penelitian sebelumnya. Penelitian ini juga berfokus pada pengembangan virtual pet dengan vuforia akan tetapi tujuan penelitian ini adalah sebuah aplikasi virtual pet yang open source.
3. Walking Your Virtual Dog: Analysis of Awareness and Proxemics with Simulated Support Animals in Augmented Reality [11]
   Penelitian ini berfokus pada binatang anjing saja dimana tujuan penelitian ini adalah mempelajari tingkah laku pengguna ketika binatang peliharaannya menghargai orang lain atau ketika orang lain menghargai binatang peliharaannya.
4. Using Augmented Realitu to Enhace Aetherpet, a Prototype of a Social Game [12]
   Penelitian ini sama dengan penelitian sebelumnya dimana peneliti berusaha mengembangkan virtual pet menggunakan augmented reality vuforia. Dan disimpulkan augmented reality dapat dikembangkan untuk virtual pet.

III. DESAIN SISTEM GAME

Pada bagian ini dijelaskan mengenai arsitektur sistem dari game, alur permainan, serta desain interface yang digunakan pada penelitian ini.

A. *Desain Arsitektur Game*



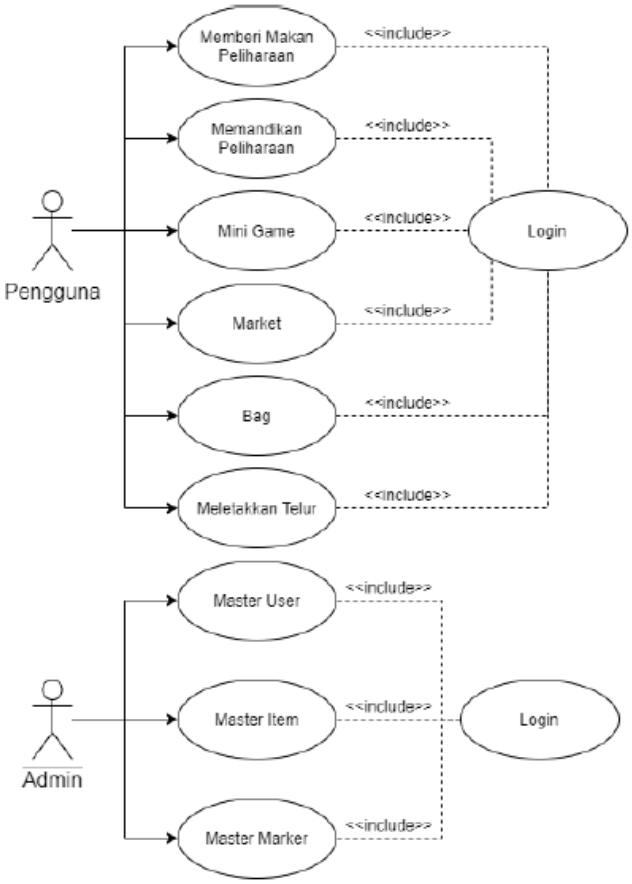
Gambar. 1. Desain Arsitektur Sistem Game

Pada bagian ini dijelaskan mengenai arsitektur sistem (Gambar 1) yang kami gunakan. Desain arsitektur ini menggambarkan gambaran secara sistematis aplikasi pada penelitian ini dimana penggambarannya didasarkan masing-masing role user yang terkait. Terdapat 2 role dalam game ini, yaitu pengguna (user) dan admin. Pengguna dapat membuat akun dengan register, login dan mendapat data-data pengguna seperti username pengguna, nama peliharaan dan jumlah koin yang dimiliki pengguna tersebut. Pengguna juga dapat meletakkan peliharaan dengan melakukan scan pada bidang datar kemudian peliharaan dan objek 3D mini gamenya akan diletakkan (Fitur ini dibuat menggunakan teknologi yang dimiliki oleh AR Foundation yaitu Plane Detection). Pengguna juga dapat bermain mini game dengan menekan objek 3D yang berada di samping peliharaan, di setiap mini game nya juga terdapat menu highscore yang digunakan untuk melihat 10 score tertinggi dan score

pengguna, setelah pengguna selesai bermain mini game maka pengguna akan diberi koin dan peliharaannya akan diberi experience. Pengguna juga dapat membeli item di market dan item yang dibeli dari market tersebut akan masuk ke bag. Sedangkan role Admin dapat melihat, menghapus dan melakukan ban pada pengguna yang melakukan kecurangan. Admin dapat melihat, menambahkan, mengubah dan menghapus data item makanan yang ada pada market. Dan terakhir, Admin juga dapat melihat, menambahkan, mengubah, menghapus dan memilih marker yang dipakai.

Gambar. 2. Alur Game

Gambar 2 adalah sebuah gambar alur game yang dikembangkan. Pengguna pada awalnya akan diminta untuk melakukan login atau register, setelah itu pengguna diminta untuk melakukan scan ke tempat peliharaan akan diletakkan, jika scan pada tempat yang diinginkan berhasil maka peliharaan akan diletakkan ditempat tersebut. Selanjutnya pengguna akan diberikan tampilan awal yang berupa menu dalam menu tersebut pengguna dapat memilih ransel, market ataupun keluar dari game. Jika pengguna memilih ransel maka pengguna akan diberi tampilan ransel, disana pengguna dapat melihat barang-barang yang dimiliki oleh pengguna tersebut. Jika pengguna memilih market maka pengguna akan diberi tampilan market, disana pengguna dapat membeli barang-barang seperti makanan dan telur. Jika pengguna memilih keluar, maka pengguna akan diarahkan untuk keluar dari game tersebut. Jika pengguna menekan objek 3D yang ada di samping peliharaan yang berhasil diletakkan, maka pengguna akan diarahkan ke mini game sesuai dengan mini game yang pengguna pilih, dalam mini game tersebut pengguna dapat memilih untuk bermain atau melihat HighScore, ketika mini game berakhir pengguna akan diberi tampilan score yang didapatkan. Jika pengguna menekan tombol makan dan melakukan scan pada gambar yang berisi marker, maka pada bagian atas marker tersebut akan muncul objek 3D makanan secara random yang dapat di berikan pada peliharaan, untuk mengganti makanannya pengguna harus mengarahkan kameranya sampai objek 3D makanannya tidak terlihat, maka akan dilakukan random lagi pada makanan yang keluar tersebut. Jika pengguna menekan tombol mandi dan melakukan scan pada gambar yang berisi marker, maka pada bagian atas marker tersebut akan muncul objek 3D tempat mandi yang dapat diberikan pada peliharaan.



Gambar. 3. Use Case Diagram Sistem Game

Gambar 3 merupakan gambar use case diagram sistem game ini. Terdapat 2 aktor, yaitu pengguna dan admin. Setiap aktor yang melakukan interaksi dengan sistem seperti menggunakan fiturnya harus login ke akun yang telah teregister sebelumnya, setelah itu pengguna dapat memberi makan peliharaan, memandikan peliharaan, memainkan mini game, membeli di market dan melihat bag yang berisi barang atau item yang sudah didapatkan atau dibeli di market. Sedangkan admin dapat melihat data user, melakukan ban pada user yang melakukan kecurangan dan menghapus user pada halaman master user. Admin juga dapat melihat, menambah, mengubah dan menghapus data item yang ada di database pada halaman master item. Selain itu Admin juga dapat melihat, menambah, mengubah, menghapus dan memilih marker yang akan digunakan pada halaman master marker.

B. Desain Interface Game

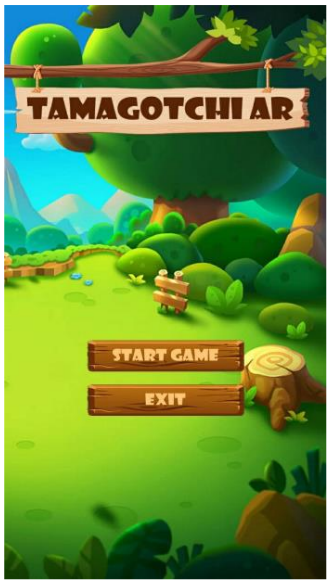
Pada bagian ini akan dijelaskan mengenai tampilan dari game yang dibuat. Bagian ini akan menunjukkan tampilan menurut sudut pandang user, tampilan ini akan terbagi menjadi dua, yaitu tampilan yang dimiliki user dan tampilan yang dimiliki oleh admin. Tampilan yang digunakan oleh user dan admin yang ditampilkan melalui mobile.

Gambar. 4. Desain User Login

Gambar 4 merupakan tampilan login. Pengguna dapat login apabila pengguna telah membuat akun dibagian registrasi. Jika pengguna belum memiliki akun, pengguna dapat menekan tombol register di kiri bawah untuk langsung diarahkan ke halaman register.

Gambar. 5. Desain User Register

Gambar 5 merupakan tampilan register pengguna. Pengguna dapat membuat akun pada halaman register ini dengan dengan mengisi 5 input yang tersedia pada halaman register pengguna dan halaman register peliharaan dan melewati pengecekan email dan password yang sesuai dengan confirm password, maka akun pengguna akan dibuat.



Gambar. 6. Desain Menu Utama

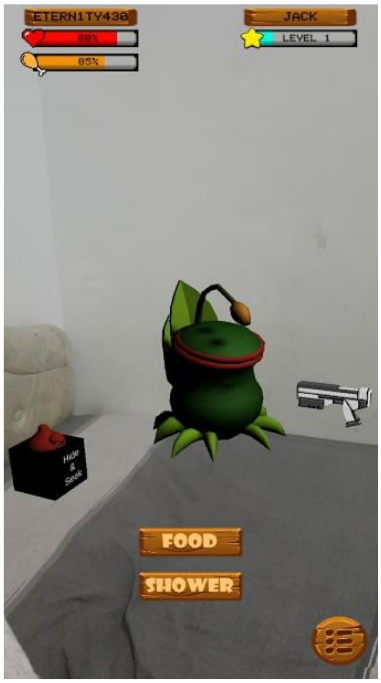
Gambar 6 merupakan tampilan menu awal game. Terdapat 2 tombol yang ada di halaman ini, yaitu tombol start game dan tombol exit. Jika pengguna menekan tombol exit maka pengguna akan diarahkan keluar dari game. Jika pengguna menekan tombol start game maka game akan dimulai dan pengguna akan diminta untuk meletakkan peliharaaannya.

Gambar. 7. Desain Market

Gambar 7 merupakan tampilan market. disini pengguna dapat melihat jumlah koinnya dan dapat membeli item yang diinginkan dengan menekan tombol buy yang ada pada item tersebut.

Gambar. 8. Desain Bag Pengguna

Gambar 8 merupakan tampilan bag pengguna. Semua item yang pernah didapat atau dibeli oleh pengguna akan disimpan pada halaman ini, disini juga pengguna dapat melihat jumlah koin yang dimilikinya dan jumlah item yang dimilikinya. Pengguna juga dapat kembali ke halaman sebelumnya dengan menekan tombol silang.



Gambar. 9. Desain Kondisi Virtual Pet

Gambar 9 merupakan tampilan dari permainan virtual pet. Pet akan otomatis muncul saat game medeteksi bidang datar. Disebelah kanan dan kiri pet adalah object 3D yang apabila ditekan akan masuk ke mini games. Terdapat 2 jenis mini games pada permainan ini. Selain mini games, pet juga memiliki status dimana apabila status berkurang harus diberi makan atau dimaindikan.

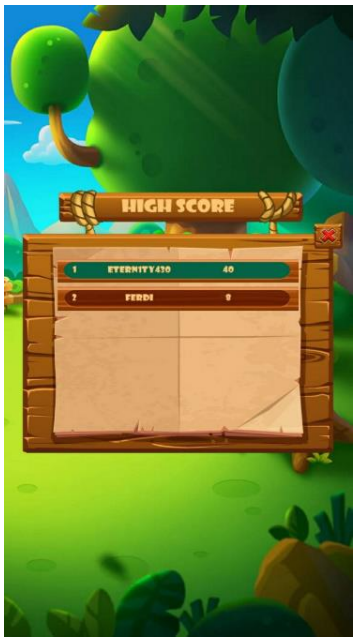
Gambar. 10. Desain Mini Game Pet Shooter

Gambar 10 merupakan tampilan dari mini game Pet Shooter. Pada mini game ini pengguna diminta untuk mencari di dunia nyata dan menembak musuh yang akan spawn terus menerus, score akan bertambah ketika musuh tersebut mati dan game ini akan berakhir ketika nyawa pengguna habis, terdapat juga tombol pause yang berada di kanan atas yang digunakan untuk melakukan pause mini game.

Gambar. 11. Desain Mini Game Hide and Seek

Gambar 11 merupakan tampilan dari minigame Hide And Seek. Pada mini game ini pengguna diminta untuk mencari peliharaannya di dunia nyata, score akan bertambah ketika

pengguna berhasil menemukan dan menekan peliharaannya, mini game ini akan berakhir ketika waktu yang ditentukan telah habis.



Gambar. 12. Desain High Score pada Mini Games

Gambar 12 merupakan tampilan dari highscore pada mini game. Pada halaman ini pengguna dapat melihat sepuluh score tertinggi dan score tertingginya. Score milik pengguna tersebut akan berwarna hijau dan sisanya akan berwarna cokelat. Jika pengguna menekan tombol silang maka pengguna akan diarahkan kembali ke halaman sebelumnya.

IV. UJI COBA

Pada tahap ini, akan dijelaskan mengenai hasil uji coba yang telah dilakukan oleh penulis. Diharapkan dengan adanya uji coba ini, penulis dapat mengetahui pendapat orang-orang terhadap game ini.

TABEL I
DAFTAR PERTANYAAN KUISIONER

No	Pertanyaan
1	Jenis Kelamin
2	Usia
3	Seberapa Sering Anda Bermain Game
4	Seberapa Sering Anda Bermain Virtual Pet
5	Aapakah Game Berjalan Lancar di Android Anda
6	Apakah AR Berjalan Lancar di Android Anda
7	Berapa Lama Anda Memainkan Game ini
8	Apakah Anda Menikmati Permainan ini
9	Apakah Anda akan Memainkan Game ini lagi jika ada update

Metode uji coba pada penelitian ini menggunakan metode kuesioner, namun sebelum mengisi kuesioner 40 responden diminta untuk melakukan uji coba pada game yang dikembangkan selama 3 bulan. Data ini dikumpulkan melalui beberapa pertanyaan yang disusun dalam bentuk kuesioner seperti disajikan pada Tabel I.

TABEL II
HASIL KUISIONER

Pertanyaan Nomor	Pilihan I	Pilihan II	Pilihan III	Pilihan IV
3	Sangat Sering **(87.5%)**	Sering **(12.5%)**	Jarang **(0%)**	Tidak Pernah **(0%)**
4	Sangat Sering **(50%)**	Sering **(20%)**	Jarang **(25%)**	Tidak Pernah **(5%)**
5	Sangat Lancar **(80%)**	Lancar **(20%)**	Kurang Lancar **(0%)**	Tidak Berjalan **(0%)**
6	Sangat Lancar **(82.5%)**	Lancar **(7.5%)**	Kurang Lancar **(7.5%)**	Tidak Berjalan **(2.5%)**
7	3 bulan **(15%)**	> 1 bulan **(52.5%)**	> 1 minggu **(25%)**	< 1 hari **(7.5%)**
8	Sangat Menikmati **(22.5%)**	Menikmati **(67.5%)**	Kurang Menikmati **(7.5%)**	Tidak Menikmati **(2.5%)**
9	Pasti **(20%)**	Sekali-sekali **(17.5%)**	Dipikir Dahulu **(62.5%)**	Tidak **(0%)**

Tabel II merupakan tabel hasil kuisioner yang telah dilakukan pada penelitian ini. Dari sini tampak bahwa responden mayoritas adalah pemain game dan tidak sedikit yang juga merupakan pemain virtual pet. Berdasarkan kuisioner tersebut secara teknis game tidak ada masalah untuk seluruh pengguna tetapi dalam upaya meningkatkan kesenangan dari seorang pemain, game ini dirasa belum cukup. Hal tersebut dapat dilihat dari pertanyaan nomor 7, dimana hanya 15% responden yang bermain penuh 3 bulan sedangkan ada 7.5% responden yang hanya bermain 1 hari saja. Berdasarkan pertanyaan nomor 8 dapat kita lihat juga bahwa terdapat 90% user yang menikmati permainan ini akan tetapi masih ada 10% user yang tidak menikmatinya. Sedangkan berdasarkan pertanyaan nomor 9 mayoritas user berpikir dahulu untuk lanjut bermain apabila terdapat update pada permainan. Melihat hal ini, penulis melakukan wawancara singkat kapada beberapa responden yang tidak menikmati game ini dan ditemukan bahwa responden tersebut merasa bosan karena diulang terus, jumlah monster yang diternakan kurang banyak, dan kurangnya fitur terkait monster itu sendiri misalnya seperti alur evolusi dan battle antar monster.

V. KESIMPULAN

Pada bagian ini dijelaskan kesimpulan yang didapat selama pembuatan penelitian dimana kesimpulan ini diperoleh dari hasil uji coba yang telah dilakukan. Kesimpulan tersebut adalah sebagai berikut:

1) Game Virtual Pet secara teknis tidak mengalami kendala bahkan dapat berjalan lancar pada berbagai mobile device yang dibuktikan pada kuisioner nomor 5 dan 6 yang menyatakan bahwa 90% lebih permainan dan AR berjalan lancar.
2) Game Virtual Pet mudah ditinggalkan oleh pemain dikarenakan sifat permainan yang berulang-ulang, hal ini dibuktikan melalui kuisioner nomor 7 dimana terdapat 32.5% pemain yang berhenti sebelum 1 bulan.
3) Game Virtual Pet kurang meningkatkan kesenangan pemain dalam jangka waktu lama hal ini dapat

dibuktikan pada pertanyaan nomor 8 yang mayoritas menikmati akan tetapi pada pertanyaan nomor 7 terdapat 35% pemain yang berhenti bermain sebelum 1 bulan.

DAFTAR PUSTAKA

[1] J. K. Haas, "A history of the unity game engine," *Diss. WORCESTER Polytech. Inst.*, p. 32, 2014.

[2] Anonymous, "Firebase," 2016. https://firebase.google.com/.

[3] C. Wicaksana, "Studi Pengkajian ARCore untuk Pengembangan Aplikasi Augmented Reality pada Android.," 2020.

[4] A. Nugroho and B. A. Pramono, "Aplikasi Mobile Augmented Reality Berbasis Vuforia Dan Unity Pada Pengenalan Objek 3D Dengan Studi Kasus Gedung M Universitas Semarang," *J. Transform.*, vol. 14, no. 2, pp. 86–91, 2017.

[5] Sinarta, "Aplikasi Augmented Reality Berbasis Android untuk Salesman Canvasing," 2017.

[6] N. Wahyudi, R. A. Harianto, and E. Setyati, "Augmented Reality Marker Based Tracking Visualisasi Drawing 2D ke dalam Bentuk 3D dengan Metode FAST Corner Detection," *J. Intell. Syst. Comput.*, vol. 1, no. 1, pp. 9–18, 2019.

[7] E. Ardhianto, W. Hadikurniawati, and E. Winarno, "Augmented Reality Objek 3 Dimensi dengan Perangkat Artoolkit dan Blender," *J. Teknol. Inf. Din. Vol.*, vol. 17, no. 2, pp. 107–117, 2012.

[8] N. Nasruddin, H. Azis, and D. Lantara, "Pengenalan Jenis Laptop menggunakan Metode Markerless," in *Prosiding SAKTI (Seminar Ilmu Komputer dan Teknologi Informasi)*, 2018, vol. 3, no. 2, pp. 148–151.

[9] J. Chahyana and V. Yesmaya, "Virtual Pet Simulator Game Using Augmented Reality on Android Platform," in *Journal of Physics: Conference Series*, 2020, vol. 1566, no. 1, p. 12088.

[10] A. Costa, R. Lima, and S. Tamayo, "Eva: a virtual pet in augmented reality," in *2019 21st Symposium on Virtual and Augmented Reality (SVR)*, 2019, pp. 47–51.

[11] N. Norouzi *et al.*, "Walking your virtual dog: Analysis of awareness and proxemics with simulated support animals in augmented reality," in *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2019, pp. 157–168.

[12] F. Kwik and R. Bahana, "Using augmented reality to enhance aetherpet, a prototype of a social game," *Procedia Comput. Sci.*, vol. 59, pp. 282–290, 2015.

INSYST

Journal of Intelligent System and Computation

Volume 03 Nomor 02 Oktober 2021

Author Guidelines

- Manuscript should be written in Indonesia and be submitted online via journal website. Online Submission will be charged at no Cost
- Manuscript should not exceed 15 pages including embedded figures and tables, without any appendix, and the file should be in Microsoft Office (.doc/.docx). download template
- Title, Abstract and Keywords must be written in bilingual
- Title should be less than 15 words
- Abstracts consists of no more than 200 words, contains the essence of the article and includes a brief background, objectives, methods and results or findings of the study. Abstract is written in one paragraph.
- Keywords are written in Indonesia and English three to five words/phrases, separated with coma and consist of important words/phrases from the article.
- Author's name, affiliation, affiliation address and email. State clearly and include country's name on your affiliation address.
- The main text of the writing should be consists of: Introduction, Method, Result and Discussion, and Conclusion; followed by Acknowledgment and Reference
- Introduction State adequate background, issues and objectives, avoiding a detailed literature survey or a summary of the results. Explain how you addressed the problem and clearly state the aims of your study.
- Used method is the scientific in the form of study of literature, observation, surveys, interviews, Focus Group Discussion, system testing or simulation and other techniques commonly used in the world of research. It is also recommended to describe analysis techniques used briefly and clearly, so that the reader can easily understand.
- Results should be clear, concise and not in the form of raw data. Discussion should explore the significance of the results of the work, not repeat them. Avoid extensive citations and discussion of published literature. INSYST will do the final formatting of your paper.
- Conclusion should lead the reader to important matter of the paper. Authors are allowed to include suggestion or recommendation in this section. Write conclusion, suggestion and/or recommendation in narrative form (avoid of using bulleting and numbering)
- Acknowledgments. It is highly recommended to acknowledge a person and/or organizations helping author(s) in many ways. Sponsor and financial support acknowledgments should be included in this section. Should you have lots of parties

to be acknowledged, state your acknowledgments only in one paragraph. Avoid of using bulleting and numbering in this section

- The number of references are not less than 10 with at least 8 primary references. Primary references are include journal, thesis, disertation and all kinds of research reports. All refferences must come from source published in last 7 years.
- Figure and table should be in black and white, and if it is made in color, it should be readable when it is later printed in black and white.
- Figure and table should be clearly readable and in a proportional measure to the overall page.

Tim Redaksi

Journal of Intelligent System and Computation

Departement of Informatics

Institut Sains dan Teknologi Terpadu Surabaya

Jl. Ngagel Jaya Tengah 73-77 Surabaya

Email: insyst@istts.ac.id

Website: https://jurnal.stts.edu/index.php/INSYST/index