****

**OTOMATISASI EKSTRAKSI KATA KUNCI PADA *WEBSITE* IMAJI SOCIOPRENEUR MENGGUNAKAN METODE *TEXTRANK*.**

**SKRIPSI**

Oleh:

**Mokhamad Asif**

**NIM 172410102039**

**PROGRAM STUDI TEKNOLOGI INFORMASI**

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS JEMBER**

**2021**

# Daftar Isi

[Daftar Isi ii](#_Toc70742412)

[Daftar Tabel iv](#_Toc70742413)

[Daftar Gambar v](#_Toc70742414)

[A. Judul 1](#_Toc70742415)

[B. Latar Belakang 1](#_Toc70742416)

[C. Rumusan Masalah 2](#_Toc70742417)

[D. Tujuan & Manfaat 2](#_Toc70742418)

[D.1 Tujuan 2](#_Toc70742419)

[D.2 Manfaat 2](#_Toc70742420)

[E. Batasan Masalah 3](#_Toc70742421)

[F. Tinjauan Pustaka 4](#_Toc70742422)

[F.1 Penlitian Terdahulu 4](#_Toc70742423)

[F.2 *Website* Imaji Sociopreneur 8](#_Toc70742424)

[F.3 *Natural Language Processing* 9](#_Toc70742425)

[F.4 Kata Kunci 9](#_Toc70742426)

[F.5 *Text Preprocessing* 10](#_Toc70742427)

[F.5.1 *Case Folding* 10](#_Toc70742428)

[F.5.2 *Tokenizing* 10](#_Toc70742429)

[F.5.3 *Filtering* 11](#_Toc70742430)

[F.5.4 *Stemming* 11](#_Toc70742431)

[F.5.5 *Parts-of-Speech Tagging* 11](#_Toc70742432)

[F.6 *Textrank* 12](#_Toc70742433)

[G. Metodologi Penelitian 15](#_Toc70742434)

[G.1 Jenis Penelitian 15](#_Toc70742435)

[G.2 Objek Penelitian 15](#_Toc70742436)

[G.3 Tempat dan Waktu Penelitian 16](#_Toc70742437)

[G.4 Gambaran Sistem 16](#_Toc70742438)

[G.5 Tahapan Penlitian 17](#_Toc70742439)

[G.5.1 Pengumpulan dan pengolahan data 17](#_Toc70742440)

[G.5.2 Pembangunan REST API 18](#_Toc70742441)

[G.5.3 Penghubungan REST API Textrank yang telah dibangun ke *Website* Imaji Sociopreneur 23](#_Toc70742442)

[H. LUARAN YANG DIHARAPKAN 23](#_Toc70742443)

[I. JADWAL KEGIATAN 23](#_Toc70742444)

[Daftar Pustaka 25](#_Toc70742445)

# Daftar Tabel

[Tabel 1. Penelitian Terdahulu 4](#_Toc70742336)

[Tabel 2. Hasil *Case Folding* 18](#_Toc70742337)

[Tabel 3. Hasil *Tokenizing* 18](#_Toc70742338)

[Tabel 4. Hasil *Stemming* 19](#_Toc70742339)

[Tabel 5. Hasil *Filtering* 19](#_Toc70742340)

[Tabel 6. Hasil *POS-Tagging* 21](#_Toc70742341)

[Tabel 7. Jadwal Kegiatan 23](#_Toc70742342)

# Daftar Gambar

[Gambar 1. Tahapan Penelitian 17](#_Toc70742395)

[Gambar 2. Pembentukan *corpus POS-Tagging* Bahasa Indonesia 21](#_Toc70742396)

# A. Judul

Otomatisasi Ekstraksi Kata Kunci Pada *Website* Imaji Sociopreneur Menggunakan Metode *Textrank*.

# B. Latar Belakang

Dalam era digital ini data sangat penting dan melimpah.  Menurut KBBI, data adalah keterangan atau bahan nyata yang dapat dijadikan dasar kajian (analisis atau kesimpulan) (Kemdikbud, 2016). Data yang ada dapat digunakan untuk berbagai keperluan penelitian maupun analisis yang dapat berguna untuk perkembangan teknologi, pengambilan keputusan bagi pemerintah maupun *stakeholder* lainnya. Data dapat ditemukan dalam berbagai media seperti buku, jurnal, internet, dan lain sebagainya.

Terdapat dua jenis data yaitu data terstruktur dan data tidak terstruktur. Data yang tersebar melalui internet banyak yang tersusun rapi dan terstruktur dalam relasi *database* maupun berupa *RESTful API* yang siap digunakan hingga data yang tidak terstruktur atau *unstructured* data. Data tidak terstruktur dapat berupa gambar, musik, video maupun teks dokumen (Wongchaisuwat, 2019). Data teks termasuk ke dalam bahasa natural atau bahasa manusia sehingga sebelum digunakan untuk keperluan penelitian maupun analisis diperlukan pemrosesan lebih lanjut menggunakan kecerdasan buatan *Natural Language Processing* (*NLP*).

Imaji Sociopreneur adalah startup sosial yang berjalan di pemberdayaan masyarakat dan teknologi pertanian. Salah satu kendala yang dihadapi oleh Imaji Sociopreneur yaitu belum optimalnya pemodelan *SEO* dari *Website* Imaji Sociopreneur seperti tidak adanya *tag*, kategori, ataupun kata kunci terkait konten-konten didalam *Website* Imaji Sociopreneur. Oleh karena itu, peneliti hendak mengimplementasikan *Natural Language Processing* (*NLP*) dengan algoritma *Textrank* untuk otomatisasi ekstraksi kata kunci pada *Website* Imaji Sociopreneur sehingga dapat meningkatkan urutan kepentingan di dalam algoritma *Pagerank*.

*Natural Language Processing* (*NLP*) adalah cabang dari ilmu Kecerdasan buatan yang berfokus pada pengolahan bahasa natural atau bahasa yang digunakan manusia (Forestiero & Papuzzo, 2020). Salah satu algoritma *Natural Language Processing* (*NLP*) yang digunakan dalam ekstraksi data tekstual adalah *Textrank*. *Textrank* adalah algoritma dengan pemodelan berbasis graf ranking model yang dibangun atas algoritma *Pagerank* yang terkenal. *Pagerank* sendiri adalah algoritma untuk menentukan pentingnya halaman *Website* berdasarkan informasi referensinya. Intinya, *Pagerank* akan melakukan pemeringkatan berbasis graf yang menghitung skor kepentingan setiap *node* atau *vertex* berdasarkan informasi dari keseluruhan graf. *Textrank* mengadopsi pemodelan pemeringkatan kepentingan setiap *node* yang akan dibangun sebuah graf sesuai dengan *Pagerank* yang diimplementasi pada data tekstual (Wongchaisuwat, 2019).

# C. Rumusan Masalah

Berdasarkan uraian yang telah disampaikan dalam latar belakang maka permasalahan yang harus diselesaikan dalam penelitian ini sebagai berikut:

1. Bagaimana cara menerapkan algoritma *Textrank* untuk penentuan *tag*, kategori atau kata kunci pada *Website* Imaji Sociopreneur?

# D. Tujuan & Manfaat

## D.1 Tujuan

Tujuan yang ingin dicapai dalam penelitian ini adalah:

1. Untuk merancang ulang sistem pada *Website* Imaji Sociopreneur pada bagian penentuan *tag*, kategori atau kata kunci dengan menggunakan algoritma *Textrank*

## D.2 Manfaat

Manfaat penelitian ini diharapkan dapat memberi manfaat sebagai berikut:

1. Bagi Peneliti

Menerapkan ilmu pengetahuan yang telah didapat selama masa perkuliahan di Fakultas Ilmu Komputer untuk mengembangkan *Website* Imaji Sociopreneur menggunakan algoritma *Textrank*.

1. Bagi Objek Penelitian

Hasil penelitian ini diharapkan dapat membantu Imaji Sociopreneur dalam peningkatan kualitas *Website* Imaji Sociopreneur pada bagian otomatisasi kata kunci.

1. Bagi Akademis

Hasil penelitian ini diharapkan dapat menambah informasi yang berkaitan dengan judul penelitian bagi peneliti lain terutama pada Fakultas Ilmu Komputer Universitas Jember. Selain itu penelitian ini bertujuan untuk mendorong peneliti lain khususnya di Fakultas Ilmu Komputer Universitas Jember untuk mendorong pembaca dengan minat yang sama untuk mengambil topik penelitian yang serupa.

# E. Batasan Masalah

Beberapa hal yang menjadi batasan masalah dalam penelitian ini sebagai berikut.

1. *Website* Imaji Sociopreneur yang dikembangkan adalah *Website* official dari Imaji Sociopreneur yang sudah digunakan hingga sekarang.
2. Pengembangan yang dilakukan pada penelitian ini hanya pengembangan otomatisasi ekstraksi kata kunci pada halaman-halaman *Website* Imaji Sociopreneur pada fitur blog, project dan event.
3. Responden dalam penelitian ini adalah jajaran dari Imaji Sociopreneur terutama pada yang menulis konten di *Website* Imaji Sociopreneur.

# F. Tinjauan Pustaka

## F.1 Penlitian Terdahulu

Tabel 1. Penelitian Terdahulu

|  |  |  |
| --- | --- | --- |
| A | Nama | Papis Wongchaisuwat |
| Tahun | 2019 |
| Judul | *Automatic Keyword Extraction Using TextRank* |
| Analisis | Kualitatif |
| Hasil Penelitian | Algoritma yang diusulkan bertujuan untuk memproses dan mengekstrak kata kunci secara otomatis dari data tekstual yang panjang. Ini terutama didasarkan pada algoritma *Texrank* dengan penyesuaian lebih lanjut pada bobot *edge* dalam graf yang dibangun dari teks input. Secara khusus, setiap *node* mewakili unit teks dasar sementara setiap *edge* diberi bobot dengan ukuran kesamaan *Word2Vec* yang disesuaikan dengan skor pentingnya kalimat. Performa superior dari algoritme yang diusulkan mungkin dapat dicapai jika bobot *edge* lebih mewakili hubungan sebenarnya di antara unit teks yang berdekatan. Bobot *edge* didasarkan terutama pada skor kepentingan kalimat dan kata. Karenanya, kesamaan kalimat yang lebih akurat berpotensi meningkatkan performa algoritme secara keseluruhan. Mempertimbangkan pengetahuan khusus domain tambahan kemungkinan akan memungkinkan algoritme yang diusulkan untuk mengekstrak kata kunci yang lebih berwawasan. Pekerjaan masa depan lainnya termasuk menerapkan algoritma yang diusulkan ke beberapa kumpulan data untuk memverifikasi generalisasi algoritma di berbagai bidang. |
| B | Nama | Rifa Khoirunisa, Erwin Apriliyanto, Arif Setia Sandi A, Kusrini |
| Tahun | 2020 |
| Judul | Penggunaan *Natural Language Processing* Pada Chatbot Untuk Media Informasi Pertanian |
| Analisis | Kualitatif |
| Hasil Penelitian | Aplikasi ini mampu mempharsing kalimat sebanyak 86,12% dari pertanyaan yang diajukan, menjawab relevansi waktu sebanyak rata-rata 70% sesuai dengan pertanyaan yang diajukan oleh user dan juga rata-rata kemampuan menampilkan jawaban yang sesuai dengan pertanyaan sebanyak 73,33%. Dari hasil penelitian ini belum ada algoritma yang digunakan untuk menampung kata yang tidak terdeteksi dan kemudian kata tersebut dapat langsung ditambahkan kedalam *database* kata dasar. Hal ini dapat dikembangkan penelitian selanjutnya. |
| C | Nama | Zhou Qingyun, Fang Yuansheng, Shang Zhenlei, Zhong Wanli |
| Tahun | 2020 |
| Judul | *Keyword Extraction Method for Complex Nodes*  *Based on TextRank Algorithm* |
| Analisis | Kualitatif |
| Hasil Penelitian | Algoritma TextRank ditingkatkan, dan metode ekstraksi kata kunci Cina berdasarkan penghapusan *node* jaringan kompleks diusulkan dalam makalah ini, yang meningkatkan efektivitas algoritma ekstraksi kata kunci Cina. Sementara itu, jumlah iterasi TextRank berkurang selama operasi aktual, dan waktu untuk ekstraksi kata kunci teks juga dipersingkat. Selain itu, hasil percobaan menunjukkan bahwa untuk algoritma ekstraksi kata kunci tanpa pengawasan, pertimbangan penuh pada karakteristik jaringan yang kompleks dari *node* kata dalam dokumen itu sendiri dapat mempengaruhi pentingnya kata dalam teks sampai batas tertentu. Oleh karena itu, metode penghapusan *node* dalam jaringan yang kompleks dapat mengekstrak kata kunci dengan lebih baik. |
| D | Nama | Nisa Kurniasih Wangsanegara, Beki Subaeki |
| Tahun | 2015 |
| Judul | Implementasi *Natural Language Processing* Dalam Pengukuran Ketepatan Ejaan Yang Disempurnakan (EYD) Pada Abstrak Skripsi Menggunakan Algoritma Fuzzy Logic |
| Analisis | Kuantitatif |
| Hasil Penelitian | Berdasarkan uraian implementasi dan pengujian aplikasi pengukur ketepatan penggunaan EYD dapat disimpulkan bahwa metode Fuzzy Logic Tsukamoto dapat diterapkan dalam perhitungan nilai penggunaan tanda baca dan penulisan kata yang terdapat pada sebuah abstrak skripsi. Hasil pengujian terhadap 20 abstrak skripsi menunjukkan 70% sesuai dengan pemeriksaan manual. |
| E | Nama | Eris , Viny Christanti M , Jeanny Pragantha |
| Tahun | 2017 |
| Judul | Penerapan Algoritma Textrank Untuk Automatic Summarization Pada Dokumen Berbahasa Indonesia |
| Analisis | Kuantitatif |
| Hasil Penelitian | Setelah melakukan percobaan, maka dapat disimpulkan bahwa:   1. Algortima TextRank dapat mengambil kalimat menjadi hasil ringkasan jika kalimat tersebut mempunyai nilai content overlap similarity yang tinggi dibandingkan dengan kalimat-kalimat yang lainnya sehingga kalimat yang direpresentasikan sebagai *vertex* tersebut mempunyai banyak *edge* dan bernilai tinggi. 2. Hasil ringkasan dengan panjang 50% mampu menghasilkan informasi sebesar 82.48% sehingga dengan membaca setengah dari dokumen asli pengguna dapat mengetahui informasi yang terkandung dari dokumen tersebut sebanyak 82.48%. Hal ini tentu dapat menghemat banyak waktu pengguna dalam membaca suatu dokumen. 3. Hasil ringkasan dengan panjang 75% dapat menghasilkan informasi sebesar 93.76%, yang berarti hampir mendekati informasi keseluruhan yang terkandung dalam dokumen asli tersebut. |
| F | Nama | Nisa Kurniasih Wangsanegara, Beki Subaeki |
| Tahun | 2015 |
| Judul | Implementasi *Natural Language Processing* Dalam Pengukuran Ketepatan Ejaan Yang Disempurnakan (EYD) Pada Abstrak Skripsi Menggunakan Algoritma Fuzzy Logic |
| Analisis | Kuantatif |
| Hasil Penelitian | Berdasarkan uraian implementasi dan pengujian aplikasi pengukur ketepatan penggunaan EYD dapat disimpulkan bahwa metode Fuzzy Logic Tsukamoto dapat diterapkan dalam perhitungan nilai penggunaan tanda baca dan penulisan kata yang terdapat pada sebuah abstrak skripsi. Hasil pengujian terhadap 20 abstrak skripsi menunjukkan 70% sesuai dengan pemeriksaan manual. |

## F.2 *Website* Imaji Sociopreneur

*Website* Imaji Sociopreneur merupakan *Website* official Imaji Sociopreneur yang berfungsi sebagai bentuk media exposure di tingkat yang lebih luas dapat memudahkan masyarakat luas untuk mencari informasi tentang Imaji Sociopreneur. Adapun fitur yang telah ada pada *Website* Imaji Sociopreneur yaitu diantaranya :

1. Beranda atau home berisi tentang profil singkat,  partner Imaji Sociopreneur, blog terbaru, event terbaru, project terbaru.
2. Blog berisi tentang konten-konten tekstual yang ditulis oleh pihak Imaji Sociopreneur maupun pihak luar yang masih terhubung dengan Imaji Sociopreneur.
3. Event berisi tentang event yang akan dijalankan, sedang dijalankan maupun telah dijalankan oleh Imaji Sociopreneur.
4. Project berisi tentang project yang akan dijalankan, sedang dijalankan maupun telah dijalankan oleh Imaji Sociopreneur.
5. Tentang berisi tentang profil lengkap dari Imaji Sociopreneur mulai dari sejarah, visi misi hingga tim dari Imaji Sociopreneur.
6. Tanya Jawab berisi tentang jawaban dari pertanyaan-pertanyaan umum tentang Imaji Sociopreneur

*Website* Imaji Sociopreneur ini masih dalam tahap perkembangan sehingga beberapa fitur masih belum berjalan dengan sempurna.

## F.3 *Natural Language Processing*

*Natural Language Processing* (*NLP*) merupakan salah satu cabang ilmu AI yang berfokus pada pengolahan bahasa natural. Bahasa natural adalah bahasa yang secara umum digunakan oleh manusia dalam berkomunikasi satu sama lain. Dikatakan Nisa Kurniasih Wangsanegara dalam jurnalnya yang berjudul Implementasi *Natural Language Processing* Dalam Pengukuran Ketepatan Ejaan Yang Disempurnakan (Eyd) Pada Abstrak Skripsi Menggunakan Algoritma Fuzzy Logic dikatakan *Natural Language Processing* (Pemrograman Bahasa Alami) adalah pembuatan program yang memiliki kemampuan untuk memahami bahasa manusia. Pada prinsipnya bahasa alami adalah suatu bentuk representasi dari suatu pesan yang ingin dikomunikasikan antar manusia (Wangsanegara & Subaeki, 2015), Dikatakan pula oleh Rifa Khoirunisa dalam Jurnalnya yang berjudul Penggunaan *Natural Processing* Pada *Chatbot* Untuk Media Informasi Pertanian dikatakan *Natural Language Processing* adalah menganalisis teks dengan cara terkomputerisasi (Khoirunisa, 2020), pada dasarnya pendekatan *NLP* mampu menangkap konteks semantik yang merepresentasikan dokumen yang padat yaitu teks tanam (Forestiero & Papuzzo, 2020) dimana teks yang telah terinput akan dihitung secara matematis maupun dipisahkan kata perkata yang nantinya akan diproses ulang, karena bahasa yang diterima oleh komputer butuh untuk diproses dan dipahami terlebih dahulu supaya maksud dari pengguna bisa dipahami dengan baik oleh komputer.

*Natural Language Processing* (Pemrograman Bahasa Alami) adalah pembuatan program yang memiliki kemampuan untuk memahami bahasa manusia. Pada prinsipnya bahasa alami adalah suatu bentuk representasi dari suatu pesan yang ingin dikomunikasikan antar manusia

## F.4 Kata Kunci

Kata kunci adalah Kata atau frasa penting dalam judul, judul subjek (pendeskripsian), catatan konten, abstrak, atau teks catatan dalam katalog online atau *database* bibliografi yang dapat digunakan sebagai istilah pencarian dalam pencarian teks bebas untuk mengambil semua catatan yang berisi itu (Reitz, 2020).

Kata kunci adalah cara penting bagi orang untuk memahami dengan cepat konten dokumen dan subjek pemahaman, dan teknologi ekstraksi kata kunci adalah cara yang signifikan untuk mendapatkan makna inti dari informasi teks dengan cepat, yang memiliki berbagai aplikasi di berbagai bidang seperti intelijen, jurnalisme, pencarian informasi, dan alam. pemahaman Bahasa (Qingyun et al., 2020).

## F.5 *Text Preprocessing*

*Text Preprocessing* adalah tahapan dimana aplikasi melakukan seleksi data yang akan diproses pada setiap dokumen. Proses preprocessing ini meliputi ***Case Folding*, *Tokenizing*, *Filtering*, *Stemming*, dan** *Parts-of-Speech Tagging* (INFORMATIKALOGI, 2016).

### F.5.1 *Case Folding*

Tahap *Case folding* merupakan tahapan awal dari proses *text preprocessing* yaitu merubah semua karakter yang ada pada data menjadi huruf kecil (*lower case*)**.** Tidak semua dokumen teks konsisten dalam penggunaan huruf kapital. Oleh karena itu, peran *Case Folding* dibutuhkan dalam mengkonversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar (biasanya huruf kecil atau lowercase). Sebagai contoh, user yang ingin mendapatkan informasi “KOMPUTER” dan mengetik “KOMPOTER”, “KomPUter”, atau “komputer”, tetap diberikan hasil retrieval yang sama yakni “komputer”. *Case folding* adalah mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf ‘a’ sampai dengan ‘z’ yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter (INFORMATIKALOGI, 2016).

### F.5.2 *****Tokenizing*****

Tahap ***Tokenizing*** adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya. Tokenisasi secara garis besar memecah sekumpulan karakter dalam suatu teks ke dalam satuan kata, bagaimana membedakan karakter-karakter tertentu yang dapat diperlakukan sebagai pemisah kata atau bukan. Sebagai contoh karakter *whitespace*, seperti *enter*, *tabulasi*, *spasi* yang dianggap sebagai pemisah kata. Namun untuk karakter petik tunggal **(‘)**, titik **(.)**, semikolon **(;)**, titk dua **(:)** atau lainnya, dapat memiliki peran yang cukup banyak sebagai pemisah kata (INFORMATIKALOGI, 2016).

### F.5.3 *****Filtering*****

Tahap *Filtering* adalah tahap mengambil kata-kata penting dari hasil token. Bisa menggunakan algoritma *stoplist* (membuang kata kurang penting) atau *wordlist* (menyimpan kata penting). *Stoplist/stopword* adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag-of-words*. Contoh *stopwords* adalah “yang”, “dan”, “di”, “dari” dan seterusnya. Kata-kata seperti “dari”, “yang”, “di”, dan “ke” adalah beberapa contoh kata-kata yang berfrekuensi tinggi dan dapat ditemukan hampir dalam setiap dokumen (disebut sebagai *stopword*). Penghilangan *stopword* ini dapat mengurangi ukuran *index* dan waktu pemrosesan. Selain itu, juga dapat mengurangi level noise (INFORMATIKALOGI, 2016).

### F.5.4 *****Stemming*****

*Stemming* adalah proses untuk mengubah kata pada setiap kalimat ke bentuk dasar atau menghapus kata-kata imbuhan. Proses *stemming* pada teks berbahasa Indonesia berbeda dengan *stemming* pada teks berbahasa Inggris. Pada teks berbahasa Inggris, proses yang diperlukan hanya proses menghilangkan sufiks. Sedangkan pada teks berbahasa Indonesia semua kata imbuhan baik itu sufiks dan prefiks juga dihilangkan (INFORMATIKALOGI, 2016).

### F.5.5 *Parts-of-Speech Tagging*

*Part-of-speech (POS) tagging* atau secara singkat dapat ditulis sebagai *tag*ging merupakan proses pemberian penanda *POS* atau kelas sintaktik pada tiap kata di dalam corpus. Dikarenakan *tag* secara umum juga diaplikasikan pada tanda baca, maka dalam proses *tag*ging, tanda baca seperti tanda titik, tanda koma, dll perlu dipisahkan dari kata-kata. Oleh sebab itu, proses tokenisasi biasanya dilakukan sebelum *POS-Tagging*. Selain itu beberapa *preprocessing* juga dilakukan seperti pemisahan koma, tanda petik, dll dari kata serta dilakukan juga disambiguitas pada tanda baca penanda akhir kalimat seperti tanda titik dan tanda tanya agar dapat dibedakan dari tanda yang digunakan untuk singkatan (seperti contohnya: e.g. dan etc.) (Suhartono, 2019).

## F.6 *Textrank*

*Textrank* merupakan *graph-based ranking algorithm* (pemeringkatan dengan model graf), untuk pemrosesan teks *Textrank* sangatlah fleksibel karena dapat digunakan pada berbagai bahasa tanpa mengubah algoritmanya. Hal ini dikarenakan *Textrank* tidak memerlukan *data training* untuk proses pengelolaan dokumen (Mihalcea & Tarau, 2004). Terdapat dua jenis pengelolahan bahasa dalam TextRank, yaitu *TextRank for keyword extraction* (ekstraksi kata kunci) dan *TextRank for sentence extraction* (ekstraksi kalimat) (Eris et al., 2017).

Model *Textrank* menerapkan algoritma *Pagerank* yang terkenal dengan data Tekstual. Algoritma *Pagerank* berbasis graf digunakan untuk mengukur kepentingan relatif halaman situs *website* dalam kumpulan *hyperlink* (Wongchaisuwat, 2019). Menurut Mihalcea dan Tarau dalam jurnalnya yang berjudul *Textrank : Bringing Order into Texts* mengatakan  “Tujuan penggunaan dalam pemodelan graf dapat digunakan untuk pemungutan suara atau rekomendasi” (Mihalcea & Tarau, 2004). Setelah menetapkan setiap halaman situs web sebagai *node*, file algoritma menghitung skor kepentingan untuk setiap *node* dalam graf. Menghubungkan satu *node* ke *node* lainnya yang digunakan untuk pemungutan suara atau rekomendasi yang diberikan antara 2 *node* yang terhubung ini. Semakin tinggi jumlah suara, semakin signifikan *node* yang sesuai. Skor kepentingan untuk setiap *node* dihitung berdasarkan probabilitas perpindahan secara acak dari *node* yang tertarik ke *node* lain dalam graf.

Dilambangkan sebuah graf sebagai G(V,E), V merupakan himpunan *vertex* graph G dan E merupakan himpunan *edge*, dimana E merupakan subset dari V\*V. Untuk *vertex* Vi, In(Vi) merupakan himpunan *vertex* yang terhubung dan mengarah masuk ke dalam *vertex* Vi (predecessor), dan Out(Vi) merupakan himpunan *vertex* yang terhubung dan mengarah keluar *vertex* Vi (successor). Nilai Vi dinyatakan dalam persamaan

(F.1)

Dimana d merupakan nilai *damping factor* yang dapat diambil nilainya mulai dari 0 hingga 1 (Mihalcea & Tarau, 2004). Koneksi antar *node* dalam graf dapat ditangani secara berbeda dengan memasukkan kekuatan koneksi ke dalam model. Secara khusus, bobot *edge* W­ij yang sesuai dengan *node* Vi dan *node* Vj dipertimbangkan saat menghitung skor kepentingan. Rumus untuk skor tertimbang didefinisikan sebagai berikut

(F.2)

Dalam grafik ini, alur algoritma yang digunakan terdiri dari: A) Perhitungan skor kalimat; B) Perhitungan skor kata kunci. Fase pertama menerapkan algoritma *Textrank* pada level kalimat yang menghasilkan skor penting untuk setiap kalimat. Pada tahap kedua, variasi dari algoritma *Textrank* di tingkat kata diimplementasikan dan dikombinasikan dengan representasi kata yang disematkan. Skor kalimat dari frasa pertama diperhitungkan saat menghitung skor kata di frasa kedua. Akhirnya, kata-kata yang terkait dengan skor tertinggi diambil sebagai kata kunci yang diekstrak. Semua implementasi menggunakan python dengan implementasi Word2Vec dari genisme. Model Word2Vec terlatih yang digunakan dalam pekerjaan ini dilatih sebagai bagian dari kumpulan data Google Berita (Wongchaisuwat, 2019). Berikut alur *Textrank* dalam pendekatan Wongchaisuwat :

1. *Sentence Scores Computation*

Untuk menghitung skor kepentingan setiap kalimat, dokumen asli awalnya dipecah menjadi beberapa kalimat. Graf yang sepenuhnya terhubung dibangun di mana *node* dan *edge* mewakili kalimat dan skor kesamaan antara 2 *node* yang berdekatan. Graf kalimat Gs = (V, E) adalah graf yang tidak berarah dengan sekumpulan kalimat V dan sisi-sisi E. Setiap *edge* diberi bobot sesuai dengan skor kesamaan kalimat. Kesamaan skor antara 2 kalimat didasarkan pada kesamaan kata dan panjang kalimat seperti yang didefinisikan dalam. algoritma *Textrank* kemudian diimplementasikan pada Gs grafik sampai konvergensi. skor kepentingan untuk kalimat Vi didefinisikan sebagai WS(Vi) diambil dari algoritma.

1. *Keyword Scores Computation*

Bobot pada algoritma *Textrank* diterapkan pada tingkatan kata untuk mengekstraksi kata atau frase. Setelah *tokenizing* dokumen asli, graf Gw tidak langsung dibangun, sementara token (kata) dianggap sebagai *node*. Hubungan sesama antar kata ditambahkan ke *edge* yang menghubungkan antara *node* yang berdekatan. Ukuran kedekatan dari kata-kata yang terbentuk dipertimbangkan dalam hubungan ini. Secara khusus, *edge* antara 2 *node* apa pun ditambahkan jika dan hanya jika jarak antara 2 kata yang sesuai kurang dari ukuran jendela yang ditentukan sebelumnya.

Bobot *edge* Wij sebagian diperoleh dari kesamaan antara kata Vi danVj. Untuk meningkatkan kinerja algoritma *Textrank*, kemiripan semantik antara 2 kata dipertimbangkan. Untuk lebih spesifik, representasi vektor untuk setiap kata diambil dari model Word2Vec. Kesamaan antara vektor kata dihitung dan dimasukkan ke dalam rumus TextRank berbobot. *Out-Of-Vocabulary* yang dilambangkan sebagai kata-kata *OOV* adalah kata-kata yang tidak terlihat yang diamati hanya dalam set tes. Ini menyiratkan bahwa vektor kata untuk kata-kata *OOV* tidak dapat diambil dari model Word2Vec. Untuk menangani kata-kata *OOV*, nilai kesamaan yang telah ditentukan digunakan sebagai default.

Penelitian ini dibangun dengan asumsi bahwa kepentingan setiap kata diperoleh dari kata itu sendiri dan kalimat dari mana kata itu diambil. Skor kalimat juga berisi wawasan yang berguna untuk membantu meningkatkan kinerja algoritme. Menurut algoritma yang diusulkan, skor kalimat WSs yang dihitung dari algoritme *Textrank* pada langkah sebelumnya dinormalisasi ke rentang 0 dan 1. Bobot *edge* yang dihitung dari model Word2Vec selanjutnya disesuaikan dengan skor kalimat ini. Pada dasarnya, vektor kata yang sesuai dengan 2 *node* yang berdekatan ini Vi danVj diambil dari model Word2Vec. Kesamaan kosinus dihitung antara 2 vektor kata ini. Selain itu, sekumpulan skor kalimat yang sesuai dengan semua kalimat tempat Vi danVj diambil, dikumpulkan. Skor kalimat rata-rata di seluruh set ini kemudian dihitung. Bobot *edge* akhir Wij adalah perkalian dari skor kalimat rata-rata dan kesamaan kata. Terakhir, rumus *Textrank* berbobot dengan bobot *edge* akhir diulangi hingga konvergensi. Skor akhir WSw untuk setiap kata diambil.

Setelah mengurutkan skor kata terakhir dalam urutan terbalik, kata-kata yang sesuai dengan skor teratas dikumpulkan sebagai kata kunci potensial. Kata kunci potensial ini diproses pasca untuk mencari kata kunci multi-kata. Secara khusus, kata kunci potensial yang berdekatan yang terdapat dalam dokumen asli digabungkan menjadi kata kunci frase tunggal.

# G. Metodologi Penelitian

## G.1 Jenis Penelitian

Jenis penelitian ini menggunakan penelitian kuantitatif. Penelitian kuantitatif adalah metode yang digunakan untuk melakukan uji pada masalah penelitian yang berhubungan dengan data angka yang dapat di kalkulas. Wahidmurni (2017) mengatakan “Metode penelitian kuantitatif merupakan suatu cara yang digunakan untuk menjawab masalah penelitian yang berkaitan dengan data berupa angka dan program statistik. Untuk dapat menjabarkan dengan baik tentang pendekatan dan jenis penelitian, populasi dan sampel, instrumen penelitian, teknik pengumpulan data, dan analisis data dalam suatu proposal dan/atau laporan penelitian diperlukan pemahaman yang baik tentang masing-masing konsep tersebut”. Dalam penilitian ini data yang akan dikalukulasi adalah data testing tingkat keakurasian dari metode yang akan didapat dari testing crawling data beberapa jurnal yang memiliki kata kunci dan beberapa wawancara kepada penulis di *Website* Imaji Sociopreneur.

## G.2 Objek Penelitian

Objek penelitian merupakan *Website* Imaji Sociopreneur. Data yang didapat diperoleh dari penulis di *Website* Imaji Sociopreneur sebagai narasumber dan Jurnal Bahasa Indonesia yang tersebar di internet sebagai uji testing.

## G.3 Tempat dan Waktu Penelitian

Tempat dilaksanakan penelitian yaitu kantor kantor Imaji Sociopreneur yang berlokasi Kelurahan/Desa Kecamatan Tegal Gede, Kecamatan Sumbersari Kabupaten Jember. Waktu penelitian dilakukan selama empat bulan, dimulai dari bulan Maret sampai dengan bulan April 2021.

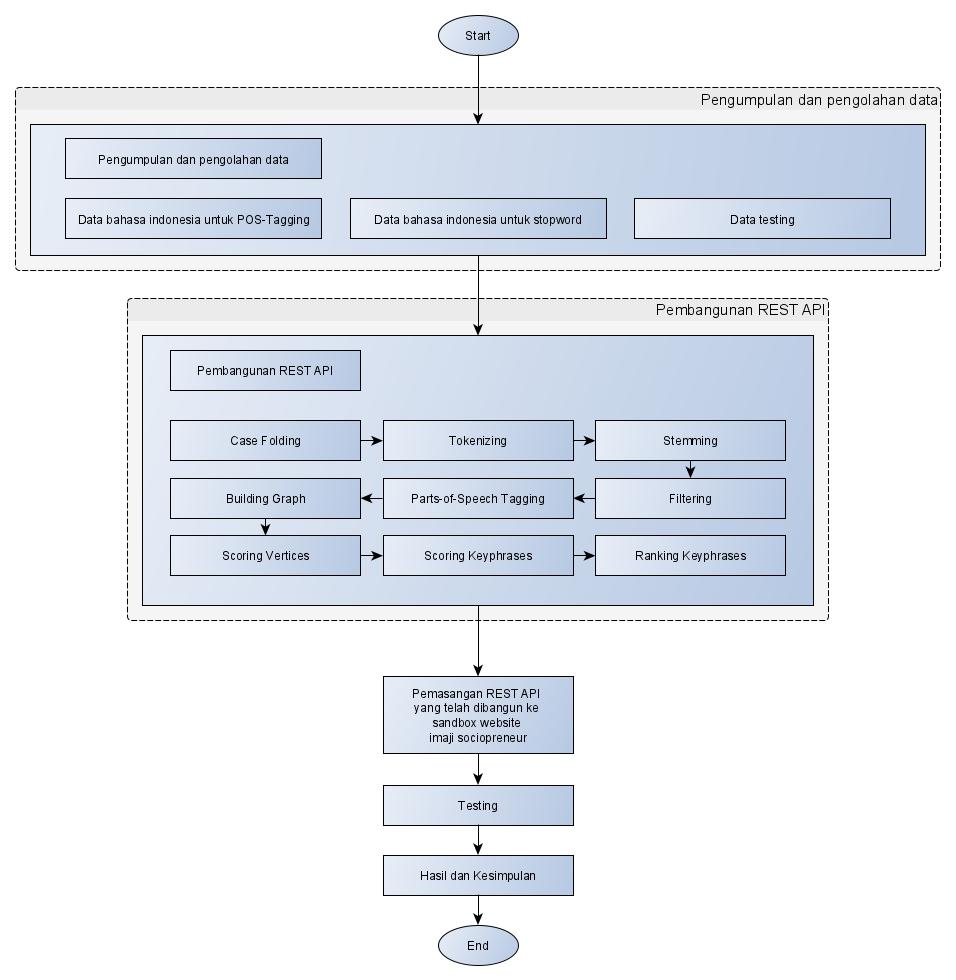
## G.4 Gambaran Sistem

Sistem akan dibangun berdasarkan pada fungsionalitas yang ada pada batasan masalah dan mengembangkan *Website* Imaji Sociopreneur yang telah ada. Data yang akan diproses oleh *Website* Imaji Sociopreneur akan dikirim melalui *RESTful API* yang dibangun khusus untuk ekstraksi kata kunci. Pembangunan *RESTful API* dilakukan agar *Website* Imaji Sociopreneur tidak terbebani oleh proses ekstraksi kata kunci yang dilakukan. Selain itu, ekstraksi kata kunci dilakukan dalam pemrosesan dengan python, sedangkan Website Imaji Sociopreneur dibangun menggunakan *Framework* Laravel.

Dalam pemrosesan pada *RESTful API* ekstraksi kata kunci setelah data diterima akan dilakukan proses mulai dari cleaning data hingga *ranking keyphrases* dari hasil *ranking keyphrases* akan dikembalikan sebagai *string* yang telah berupa *keyword* yang telah di ekstraksi.

## G.5 Tahapan Penlitian

Berikut merupakan alur dari tahapan penelitian dalam sistem :



Gambar 1. Tahapan Penelitian

### G.5.1 Pengumpulan dan pengolahan data

Pengumpulan dan pengolahan data dibagi menjadi 3 bagian. Pertama, pengumpulan dan pengolahan data bahasa Indonesia untuk *POS-Tagging* dimana kata akan diubah olah untuk didapatkan jenis-jenisnya, misalnya kata benda, kata kerja kata hubung dan lain-lain. Kedua, pengumpulan dan pengolahan data bahasa Indonesia untuk *stopword* dimana kata akan dibersihkan dari kata yang tidak diperlukan. Ketiga, data untuk testing yang dilakukan dengan cara crawling data pada website yang menyediakan hasil penelitian yang telah ada kata kuncinya yang akan dijadikan untuk mengukur keberhasilan serta data dari penulis di Website Imaji Sociopreneur.

### G.5.2 Pembangunan *RESTful API*

Dalam tahap pembangunan *RESTful API* terdapat beberapa step dalam pemrosesan teks untuk menghasilkan kata kunci yang telah didapat dari teks yang dikirim dari website imajisociopreneur maupun postman untuk testing. Dimana tahap ini melakukan

1. *Case Folding*

*Case Folding* adalah tahapan awal dari *text processing* yaitu merubah semua karakter yang ada data menjadi huruf kecil (*lowercase*). Dalam proses ini implementasinya akan menggunakan fungsi python *lower*() untuk mengubah menjadi huruf kecil. Berikut contoh hasil dari proses *case folding*:

Tabel 2. Hasil *Case Folding*

|  |  |
| --- | --- |
| **Sebelum *case folding*** | **Hasil *case folding*** |
| Melihat keadaan tersebut kami dari Imaji Sociopreneur Bersama Yayasan Mimpi Indonesia menggagas sebuah Gerakan yang kami beri nama Menanam Buku | melihat keadaan tersebut kami dari imaji sociopreneur bersama yayasan mimpi indonesia menggagas sebuah gerakan yang kami beri nama menanam buku |

1. *Tokenizing*

*Tokenizing* adalah proses pemecahan dokumen yang terdiri dari kumpulan kalimat menjadi beberapa bagian kata yang disebut token. Dalam pada implementasinya akan menggunakan fungsi *word\_tokenize(word)* dari library NLTK yang ada pada python Berikut contoh hasil dari proses *tokenizing*:

Tabel 3. Hasil *Tokenizing*

|  |  |
| --- | --- |
| **Sebelum *tokenizing*** | **Hasil *tokenizing*** |
| melihat keadaan tersebut kami dari imaji sociopreneur bersama yayasan mimpi indonesia menggagas sebuah gerakan yang kami beri nama menanam buku | ['melihat', 'keadaan', 'tersebut', 'kami', 'dari', 'imaji', 'sociopreneur', 'bersama', 'yayasan', 'mimpi', 'indonesia', 'menggagas', 'sebuah', 'gerakan', 'yang', 'kami', 'beri', 'nama', 'menanam', 'buku'] |

1. *Stemming*

*Stemming* merupakan tahapan yang diperlukan untuk memperkecil jumlah indeks dari suatu dokumen, berdasarkan kata penyusun dari dokumen tersebut. Pada proses stemming juga digunakan library Sastrawi untuk menemukan kata dasar. Berikut merupakan contoh dari proses *stemming*:

Tabel 4. Hasil *Stemming*

|  |  |
| --- | --- |
| **Sebelum *stemming*** | **Hasil *stemming*** |
| ['melihat', 'keadaan', 'tersebut', 'kami', 'dari', 'imaji', 'sociopreneur', 'bersama', 'yayasan', 'mimpi', 'indonesia', 'menggagas', 'sebuah', 'gerakan', 'yang', 'kami', 'beri', 'nama', 'menanam', 'buku'] | ['lihat', 'ada', 'sebut', 'kami', 'dari', 'imaji', 'sociopreneur', 'sama', 'yayasan', 'mimpi', 'indonesia', 'gagas', 'buah', 'gera', 'yang', 'kami', 'beri', 'nama', 'tanam', 'buku'] |

1. Filtering

*Filtering* merupakan tahapan pengambilan sejumlah kata penting dari hasil token yang telah didapatkan. Dalam hal ini tahapan algoritma yang dipakai adalah stopword, dimana data kata yang telah ada akan dicocokan dengan list stopword dan yang ada dalam kamus stopword maka data kata akan dihilangkan. Data stopword yang digunakan berasal dari <https://www.ranks.nl/stopwords/indonesian>. Berikut contoh hasil dari proses filtering:

Tabel 5. Hasil *Filtering*

|  |  |
| --- | --- |
| **Sebelum *filtering*** | **Hasil *filtering*** |
| ['lihat', **'ada'**, 'sebut', **'kami'**, **'dari**', 'imaji', 'sociopreneur', **'sama'**, 'yayasan', 'mimpi', 'indonesia', 'gagas', 'buah', 'gera', 'yang', 'kami', 'beri', 'nama', 'tanam', 'buku'] | ['lihat', 'sebut', 'imaji', 'sociopreneur', 'yayasan', 'mimpi', 'indonesia', 'gagas', 'buah', 'gera', 'beri', 'nama', 'tanam', 'buku'] |

1. *Parts-of-Speech Tagging*

*Parts-of-Speech Tagging* merupakan tahapan pemberian *tag* pada setiap *corpus*, *Parts-of-Speech Tagging* tidak hanya memberi *tag* pada kata namun juga pada symbol ataupun tanda baca. *Parts-of-Speech Tagging* yang digunakan menggunakan Flair *NLP* library yang dikembangkan oleh Puspita Kaban, Untuk melakukan *POS-tagging*, kita perlu membuat sebuah *POS-Tagger* yang terdiri atas word embedding dan dictionary. Sederhananya, *word embedding* adalah representasi dari kata-kata ke dalam sebuah vektor. Adapun library pada *tag*ger ini dibangun dari sebuah *corpus* (kumpulan kata-kata) yang sudah ditandai. Flair *NLP* sudah menyediakan corpus bahasa Indonesia yang dapat digunakan untuk *POS-*Tagging (Kaban, 2019). Namun Flair tidak menyediakan secara langsung *POS-tagging* berbahasa Indonesia maka diperlukan train library *POS-tagging*, yang dibuat dengan cara sebagai berikut



Gambar 2. Pembentukan *corpus* *POS-Tagging* Bahasa Indonesia

Maka setelah itu library flair untuk *POS-Tagging* berbahasa Indonesia dapat digunakan dan berikut contoh hasil dari *POS-Tagging*:

Tabel 6. Hasil *POS-Tagging*

|  |  |
| --- | --- |
| **Kata** | **Keterangan Label** |
| Saya | PRON |
| Dan | CCONJ |
| Dia | PRON |
| Kemarin | ADJ |
| Pergi | VERB |
| Ke | ADP |
| Pasar | NOUN |
| Bersama | ADP |
| Untuk | ADP |
| Membeli | VERB |
| Jeruk | NOUN |

Dimana keterangan label sebagai berikut:

ADJ: kata sifat

ADP: preposisi

ADV: keterangan

AUX: kata bantu

CCONJ: kata penghubung

INTJ: kata seru

NOUN: kata benda

NUM : angka

PART : partikel

PRON : kata ganti

PUNCT : tanda baca

SYM : simbol

VERB : kata kerja

X : lainnya

1. *Building Graph*

*Building Graph* adalah tahapan pertama dalam *textrank*. *Building graph* dilakukan karena *textrank* adalah model berbasis graf. Setiap kata dalam kosakata akan berfungsi sebagai simpul untuk graf. Kata-kata tersebut akan direpresentasikan di simpul oleh *index*nya dalam daftar kosakata. *Building graph* dilakukan dengan cara pemanfaatan library math dan numpy pada python untuk membantu pembangunan graf.

1. S*coring Vertices*

*Scoring Vertices* adalah tahapan node atau simpul yang telah dibuat pada tahap 6 akan di hitung menggunakan persamaan [(F.1)](#_F.6_Textrank) yang akan didapatkan nilai tiap *vertex* yang akan digunakan untuk penentuak *keyphrases* atau frasa unik.

1. *Scoring Keyphrases*

*Scoring Keyphrases* adalah tahapan menilai frasa (frasa kunci kandidat) dan membangun daftar frasa kunci dengan membuat daftar versi frasa tokenized \ kandidat-frasa kunci. Frasa dinilai dengan menambahkan skor anggotanya (kata \ unit teks yang diberi peringkat oleh algoritma graf).

1. *Ranking Keyphrases*

*Ranking Keyphrases* adalah tahapan memberi peringkat frasa kunci berdasarkan skor yang telah dihitung pada proses sebelumnya. *Ranking Keyphrases* dilakukan dengan menggunakan numpy untuk melakukan *sorting*.

### G.5.3 Penghubungan *RESTful API* Textrank yang telah dibangun ke *Website* Imaji Sociopreneur

Pemasangan *RESTful API* yang telah dibangun pada proses G.4.2 dilakukan menggunakan *request* *RESTful API* milik *Website* Imaji Sociopreneur untuk menambahkan data ke *database* tulisan yang bersangkutan.

# H. LUARAN YANG DIHARAPKAN

Luaran yang diharapkan dari penelitian ini yaitu:

1. Digunakan sebagai seminar proposal
2. Skripsi sebagai tugas akhir
3. Rekomendasi bagi objek penelitian
4. Jurnal yang dipublikasikan
5. SEO Support untuk *Website* Imaji Sociopreneur

# I. JADWAL KEGIATAN

Pengerjaan skripsi ini diperlukan beberapa tahap untuk menyelesaikan, berupa jadwal kegiatan sebagai berikut :

Tabel 7. Jadwal Kegiatan

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No | Tahapan Penelitian | Februari | | | | Maret | | | | April | | | | Mei | | | |
| 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 1 | Penyusunan dan pengajuan proposal |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2 | Seminar Proposal |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 3 | Pengumpulan dan pengolahan data |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 4 | Pembangunan REST API |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 5 | Pemakaian REST API yang telah dibuat ke sandbox *Website* imaji sociopreneur |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 6 | Testing |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 7 | Hasil dan kesimpulan |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 8 | Penulisan laporan skripsi |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

# Daftar Pustaka

Eris, Mawardi, V. C., & Pragantha, J. (2017). PENERAPAN ALGORITMA TEXTRANK UNTUK AUTOMATIC SUMMARIZATION PADA DOKUMEN BERBAHASA INDONESIA. *Jurnal Ilmu Teknik Dan Komputer*, *1*(1), 71–78. https://publikasi.mercubuana.ac.id

Forestiero, A., & Papuzzo, G. (2020). Natural language processing approach for distributed health data management. *Proceedings - 2020 28th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, PDP 2020*, 360–363. https://doi.org/10.1109/PDP50117.2020.00061

INFORMATIKALOGI. (2016). *Text Preprocessing | INFORMATIKALOGI*. Informatikalogi. https://informatikalogi.com/text-preprocessing/

Kaban, P. (2019). *POS-Tagging Bahasa Indonesia dengan Flair NLP*. Medium. https://puspitakaban.medium.com/pos-*tag*ging-bahasa-indonesia-dengan-flair-nlp-c12e45542860

Kemdikbud. (2016). *Kamus Besar Bahasa Indonesia (KBBI)*. Kementerian Pendidikan Dan Kebudayaan Republik Indonesia. https://kbbi.kemdikbud.go.id/

Khoirunisa, R. (2020). Penggunaan *Natural Language Processing* Pada Chatbot Untuk Media Informasi Pertanian. In *Indonesian Journal of Applied Informatics* (Vol. 4, Issue 2, p. 55). https://doi.org/10.20961/ijai.v4i2.38688

Mihalcea, R., & Tarau, P. (2004). TextRank : Bringing Order into Texts. *ResearchGate - Conference Paper July 2004*, *May 2014*.

Qingyun, Z., Yuansheng, F., Zhenlei, S., & Wanli, Z. (2020). Keyword Extraction Method for Complex Nodes Based on TextRank Algorithm. *Proceedings - 2020 International Conference on Computer Engineering and Application, ICCEA 2020*, 359–363. https://doi.org/10.1109/ICCEA50009.2020.00084

Reitz, J. M. (2020). *Online Dictionary for Library and Information Science*. ABC-CLIO, LLC. http://www.abc-clio.com/ODLIS/

Suhartono, D. (2019). *Part of speech tagging*. Binus. https://socs.binus.ac.id/2019/12/31/part-of-speech-*tag*ging/

Wahidmurni. (2017). PEMAPARAN METODE PENELITIAN KUANTITATIF. In *Вестник Росздравнадзора*. UIN Maulana Malik Ibrahim Malang. http://repository.uin-malang.ac.id/1985/2/1985.pdf

Wangsanegara, N. K., & Subaeki, B. (2015). Implementasi *Natural Language Processing* Dalam Pengukuran Ketepatan Ejaan Yang Disempurnakan (Eyd) Pada Abstrak Skripsi Menggunakan Algoritma Fuzzy Logic. *Jurnal Teknik Informatika*, *8*(2). https://doi.org/10.15408/jti.v8i2.3185

Wongchaisuwat, P. (2019). Automatic Keyword Extraction Using TextRank. *2019 IEEE 6th International Conference on Industrial Engineering and Applications, ICIEA 2019*, 377–381. https://doi.org/10.1109/IEA.2019.8714976