# IDENTIFIKASI BERITA *HOAX* BERBAHASA INDONESIA MENGGUNAKAN *BIDIRECTIONAL LONG SHORT TERM MEMORY* (Bi-LSTM)

**SKRIPSI**

**NADIA FARHANI**

**181402015**

**PROGRAM STUDI S1 TEKNOLOGI INFORMASI**

**FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI**

**UNIVERSITAS SUMATERA UTARA**

**MEDAN**

**2024**

# IDENTIFIKASI BERITA *HOAX* BERBAHASA INDONESIA MENGGUNAKAN *BIDIRECTIONAL LONG SHORT TERM MEMORY* (Bi-LSTM)

SKRIPSI

Diajukan untuk melengkapi tugas dan memenuhi syarat memperoleh ijazah Sarjana Teknologi Informasi

NADIA FARHANI
181402015


PROGRAM STUDI S1 TEKNOLOGI INFORMASI
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS SUMATERA UTARA
MEDAN
2024

ii

## PERSETUJUAN

| | |
|---|---|
| Judul | : IDENTIFIKASI BERITA *HOAX* BERBAHASA INDONESIA MENGGUNAKAN *BIDIRECTIONAL LONG SHORT TERM MEMORY* (Bi-LSTM) |
| Kategori | : SKRIPSI |
| Nama | : NADIA FARHANI |
| Nomor Induk Mahasiswa | : 181402015 |
| Program Studi | : TEKNOLOGI INFORMASI |
| Fakultas | : ILMU KOMPUTER DAN TEKNOLOGI INFORMASI UNIVERSITAS SUMATERA UTARA |

Medan, 10 Januari 2024

Komisi Pembimbing

Pembimbing 1

Dr. Muhammad Anggia Muchtar S.T., MM.IT.

NIP. 198001102008011010

Pembimbing 2

Ivan Jaya S.Si., M.Kom.

NIP. 198407072015041001

Diketahui/disetujui oleh

Program Studi S1 Teknologi Informasi

Ketua,

Dedy Arisandi, ST., M.Kom.

NIP. 197908312009121002

# PERNYATAAN

IDENTIFIKASI BERITA *HOAX* BERBAHASA INDONESIA MENGGUNAKAN *BIDIRECTIONAL LONG SHORT TERM MEMORY* (Bi-LSTM)

SKRIPSI

Saya mengakui bahwa skripsi ini adalah hasil karya saya sendiri, kecuali beberapa kutipan dan ringkasan yang masing-masing telah disebutkan sumbernya.

Medan, 10 Januari 2024

Nadia Farhani

181402015

# UCAPAN TERIMA KASIH

Dengan mengucapkan terima kasih kepada Allah SWT atas karunia dan kasih sayang-Nya, penulis dapat menyelesaikan skripsi ini dengan judul **"Identifikasi Berita Hoax Berbahasa Indonesia Menggunakan Bidirectional Long Short Term Memory (Bi- LSTM)"** untuk menjadi salah satu syarat mendapatkan gelar Sarjana Komputer pada Fakultas Ilmu Komputer & Teknologi Informasi Universitas Sumatera Utara. Sehubungan hal demikian, penulis ingin mengucapkan terima kasih sebesar-besarnya kepada :

1. Ibu Dr. Maya Silvi Lydia B.Sc., M.Sc., selaku Dekan Fakultas Ilmu Komputer & Teknologi Informasi Universitas Sumatera Utara.
2. Bapak Dr. Mohammad Andri Budiman, S.T., M.Comp.Sc., M.E.M., selaku Wakil Dekan I Fakultas Ilmu Komputer & Teknologi Informasi Universitas Sumatera Utara.
3. Ibu Sarah Purnamawati, S.T., M.Sc., selaku Wakil Dekan II Fakultas Ilmu Komputer & Teknologi Informasi Universitas Sumatera Utara.
4. Bapak Romi Fadillah Rahmat, B.Comp.Sc., M.Sc., selaku Wakil Dekan III Fakultas Ilmu Komputer & Teknologi Informasi Universitas Sumatera Utara.
5. Dedy Arisandi S.T., M.Kom., selaku Ketua Program Studi S1 Fakultas Ilmu Komputer & Teknologi Informasi Universitas Sumatera Utara.
6. Bapak Dr. Muhammad Anggia Muchtar S.T., MM.IT., selaku Dosen Pembimbing I yang sudah membimbing serta memberikan yang terbaik bagi kelancaran skripsi penulis.
7. Bapak Ivan Jaya S.Si., M.Kom., selaku Dosen Pembimbing II yang sudah menyediakan waktu selama membimbing serta menyampaikan arahan dan masukan untuk kesempurnaan skripsi penulis.
8. Bapak Prof. Dr. Drs. Opim Salim Sitompul, M.Sc. dan Ibu Rossy Nurhasanah, S.Kom., M.Kom., selaku Dosen Pembanding yang telah memberikan kritik dan saran yang membangun dalam proses penyempurnaan skripsi ini.
9. Staff dan Pegawai Fakultas Ilmu Komputer & Teknologi Informasi Universitas Sumatera Utara yang telah membantu segala urusan administrasi dalam menyelesaikan skripsi.

10. Kedua orangtua penulis, Ayah Drs. Abdul Hafiz, MM., dan Ibu Dra. Dairina Yusri, MPdI., serta saudara kandung Kakak dr. Maulida Zahra, dan Adik Naura Kamilah yang sudah memberikan perhatian, nasehat, dukungan, doa, serta semangat untuk penulis. Terima kasih karena telah menjadi bagian atas perjalanan penulis hingga sekarang.
11. Teman-teman seperjuangan penulis, Shelli Athaya, Tengku Zalfa Qadriyya Munadhila, Nurhaliza Syahfitri, Karina Putri Kaban, Tiara Amalia, yang sudah berjuang bersama-sama dalam menyelesaikan tugas akhir dan menjadi teman terbaik selama perkuliahan ini.
12. Teman-teman terdekat penulis, Vania Anastasia, Nabila Meidira, Annisa Aulia, Husna Nabila Siregar, Amirah Jilan Fakhira, Dini Fakhira yang selalu menjadi pendukung dan pendengar yang baik serta mengajarkan banyak hal tentang pelajaran hidup bagi penulis.
13. Terima kasih kepada Fikri Fadhlillah yang sudah memberikan motivasi, dukungan, serta membantu penulis dalam menyelesaikan skripsi ini.
14. Penulis tidak dapat menyebutkan nama orang-orang yang terlibat baik secara langsung atau tidak langsung pada skripsi ini.

Semoga penulis dan setiap orang yang dia sayangi diberi rahmat, karunia, dan berkat oleh Allah SWT di dunia dan akhirat. Penulis menerima banyak dukungan, baik secara formal maupun tidak formal, dari saran orang tua, saudara kandung, dan teman-teman, serta doa dan motivasi. Penulis memberikan gelar sarjana ini kepada kalian. Penulis sadar bahwa skripsi ini masih jauh dari kata sempurna, jadi penulis sungguh mengharapkan kritikan serta saran untuk dapat membangun guna membantunya menjadi lebih baik di masa mendatang. Penulis juga berharap skripsi yang sederhana ini berguna bagi seluruh pembaca serta individu yang membutuhkannya.

Medan, 10 Januari 2024

Penulis

# ABSTRAK

## IDENTIFIKASI BERITA *HOAX* BERBAHASA INDONESIA MENGGUNAKAN *BIDIRECTIONAL LONG SHORT TERM MEMORY* (Bi-LSTM)

Berita *hoax* masih sedikit yang dapat diidentifikasi karena membutuhkan pengetahuan khusus, sementara orang yang memiliki kemampuan tersebut masih terbilang sedikit. Pada saat ini sistem identifikasi masih dilakukan secara manual, sehingga jika sebuah informasi semakin banyak dan beredar tentunya menjadi semakin sulit dan merepotkan. Oleh karena itu, dibutuhkan sebuah sistem otomatis yang mampu mengidentifikasi berita yang termasuk ke dalam kategori *hoax* atau *non-hoax*. Penelitian ini bertujuan untuk mengimplementasikan algoritma *Bidirectional Long Short Term Memory* (Bi-LSTM) dalam mengotomatisasi identifikasi judul berita *hoax* berbahasa Indonesia secara otomatis. Model menggunakan *word embedding* dalam merepresentasikan teks kedalam vektor. Hasil penelitian ini menunjukkan bahwa sebuah model dengan akurasi 91% dan sebuah sistem mampu dalam mengidentifikasi judul berita apakah termasuk kedalam *hoax* atau *non-hoax.*

**Kata kunci:** *Hoax*, Bi-LSTM, Berita, Bahasa Indonesia

# ABSTRACT

## *IDENTIFICATION OF HOAX NEWS IN INDONESIAN USING BIDIRECTIONAL LONG SHORT TERM MEMORY (Bi-LSTM)*

*There are still few hoax news that can be identified because it requires special knowledge, while there are still relatively few people who have this ability. Currently, the identification system is still done manually, so if more and more information is circulated, it will certainly become more difficult and troublesome. Therefore, an automatic system is needed that is able to identify news that falls into the hoax or non-hoax category. This research aims to implement the Bidirectional Long Short Term Memory (Bi-LSTM) algorithm to automatically identify hoax news titles in Indonesian. The model uses word embedding to represent text into vectors. The results of this research show that a model with accuracy 91% and a system is capable of identifying whether news headlines are hoaxes or non-hoaxes.*

***Keywords:*** *Hoax, Bi-LSTM, News, Indonesian*

# DAFTAR ISI

# DAFTAR TABEL

# DAFTAR GAMBAR

# BAB 1
# PENDAHULUAN

## 1.1 Latar Belakang

Penggunaan Internet di Indonesia semakin terus meningkat. Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) menulis bahwa jumlah pengguna aktif internet di Indonesia sudah mencapai 196,7 juta pada tahun 2019-2020. Hal ini dapat mengakibatkan penyalahgunaan dalam internet itu sendiri (Ismayanti, 2021). Ledakan informasi yang dapat dihasilkan oleh siapa saja dapat menimbulkan berita bohong atau *hoax* baik di saluran berita, media sosial, dan lain sebagainya. Berita *hoax* dibuat dan disebarluaskan dengan berbagai motif, seperti motif politik dan komersial (Al-Ash & Wibowo, 2018). Informasi *hoax* dapat membohongi siapa saja, tidak memandang latar belakang seseorang. Bahkan kalangan intelektual juga bisa tertipu dan mempercayai sebuah informasi *hoax*.

*Hoax* dibentuk untuk mengelabui orang agar menjalankan sesuatu dengan ancaman atau penipuan. *Hoax* dapat menimbulkan akibat negatif, seperti hilangnya reputasi, harta benda, bahkan nyawa. Semakin cepat berita *hoax* menyebar, semakin cepat pula dampaknya terhadap komunitas yang ada (Prasetijo dkk., 2019). Dampak yang dihasilkan oleh *hoax* merupakan dampak yang tidak langsung disadari oleh pembaca, tetapi dapat menyerang pemikiran dan mempengaruhi cara berpikir pembaca jika tidak berhati-hati (Pardede & Ibrahim, 2020). Mendeteksi berita *hoax* dapat dilakukan melalui identifikasi. Identifikasi adalah proses membangun model matematika dari sistem dinamis berdasarkan pengamatan dan pengetahuan sebelumnya (Norton, 2009).

Adapun metode yang dapat digunakan untuk melakukan identifikasi, diantaranya adalah *Bidirectional Long Short Term Memory* (Bi-LSTM). Metode Bi-LSTM adalah LSTM dua arah, yang berarti sinyal merambat mundur serta maju dalam waktu. Pada Bi-LSTM kita memberikan *input* baik dari arah kanan ke kiri maupun dari kiri ke kanan. Untuk mendukung proses identifikasi tersebut digunakan juga suatu teknik NLP. Salah satu teknik NLP adalah *Word2Vec*. *Word2Vec* digunakan untuk merepresentasikan kata yang terdistribusi dalam korpus C, dan menerima korpus teks sebagai *input* dan *output* representasi vektor untuk setiap kata.

Beberapa penelitian yang telah dilakukan sebelumnya adalah penelitian oleh Al-Ash & Wibowo (2018) yang mengidentifikasi karakteristik berita palsu pada dokumen yang menggunakan Bahasa Indonesia. Kinerja yang dicapai dari representasi mencapai 96,74%. Selanjutnya penelitian oleh Aziz (2019) yang mengidentifikasi artikel kesehatan yang masuk apakah tergolong dalam berita *hoax* atau fakta menggunakan kombinasi metode *K-Nearest Neighbor* dan *Naive Bayes*. Nilai akurasi optimum yang dihasilkan untuk klasifikasi artikel kesehatan sebesar 88%, untuk nilai presisi optimum sebesar 83%, sedangkan untuk nilai *recall* optimum sebesar 100%. Lalu penelitian oleh Rusli dkk. (2020) yang mengimplementasikan dan mengevaluasi algoritma *multilayer perceptron* dalam mengklasifikasikan artikel berita untuk mengidentifikasi berita palsu bersama dengan berbagai konfigurasi seperti model *n-gram* dan metode ekstraksi fitur. Masing-masing nilai presisi dan *recall* mencapai 0,84 dan 0,73, dan skor F1 rata-rata makro 0,82.

Adapun penelitian yang menggunakan metode *Bidirectional Long Short Term Memory* oleh Isnain dkk. (2020) yang mendeteksi ujaran kebencian atau bukan ujaran kebencian *tweet* berbahasa Indonesia dengan menggunakan metode *Bidirectional Long Short Term Memory* dan metode ekstrasi fitur *Word2Vec* dengan arsitektur *Continuous Bag-Of-Word* (CBOW). Nilai akurasi sebesar 94,66%, dengan masing-masing nilai presisi 99,08%, *recall* 93,74% dan F-*measure* 96,29%. Untuk Bi-LSTM dengan tiga *layer* memiliki akurasi 96,93%. Penambahan satu *layer* pada Bi-LSTM meningkat 2,27%. Kemudian penelitian oleh Hilmawan (2022) yang membuat sebuah model klasifikasi untuk memprediksi sarkasme pada judul berita berbahasa inggris. Pada Bi-LSTM mendapatkan akurasi validasi sebesar 82,55% dan F1 *score* sebesar 80,92%, dan LSTM mendapatkan akurasi validasi sebesar 81,90% dan F1 *score* sebesar 80,47%.

Penulis melakukan penelitian dengan menggunakan metode *Bidirectional Long Short Term Memory* (Bi-LSTM) dikarenakan dalam penelitian ini bertujuan untuk mengimplementasikan metode Bi-LSTM dalam mengotomatisasi identifikasi teks berita *hoax* berbahasa Indonesia. Selain itu berdasarkan penelitian terdahulu, Bi-LSTM memiliki nilai akurasi yang cukup tinggi yaitu sebesar 94,66% yang dilakukan oleh Isnain dkk. (2020). Jadi metode penelitian tersebut yang sesuai pada penelitian ini. Metode Bi-LSTM juga mempunyai kelebihan dan kekurangan, dimana kelebihan dari metode Bi-LSTM adalah dapat mengakses informasi selanjutnya dan informasi sebelumnya dengan memproses data dari 2 arah yaitu *forward* dan

*backward* serta semakin banyaknya data maka akan semakin meningkat performa algoritma tersebut. Sedangkan kekurangan dari metode Bi-LSTM adalah membutuhkan data yang banyak serta waktu dan biaya komputasi yang lebih tinggi dari metode penelitian yang lain.

Berdasarkan latar belakang tersebut dan penelitian-penelitian terdahulu, penulis mengajukan sebuah penelitian untuk mengidentifikasi berita *hoax* dengan judul "Identifikasi Berita *Hoax* Berbahasa Indonesia Menggunakan *Bidirectional Long Short Term Memory* (Bi-LSTM)".

## 1.2 Rumusan Masalah

Seiring dengan berjalannya perkembangan teknologi informasi, media berita ikut berkembang untuk menyajikan informasi dalam media *online*. Tetapi, dalam penyebarannya masih banyak berita yang ditemukan ialah berita *hoax* atau berita yang tidak benar. Banyak pembaca yang kurang memahami literasi sehingga seringkali lalaidengan pentingnya memvalidasi kebenaran dari sebuah berita. Berita *hoax* masih sedikit yang dapat diidentifikasi karena membutuhkan pengetahuan khusus, sementara orang yang memiliki kemampuan tersebut masih terbilang sedikit. Saat ini, sistem identifikasi dilakukan secara manual. Jika informasi semakin banyak dan tersebar, hal itu pasti akan menjadi lebih sulit dan merepotkan. Oleh karena itu, dibutuhkan sebuah sistem otomatis untuk mengidentifikasi berita *hoax* menggunakan *Bidirectional Long Short Term Memory* (Bi-LSTM).

## 1.3 Tujuan Penelitian

Tujuan pada penelitian ialah untuk mengimplementasikan algoritma Bi-LSTM dalam mengotomatisasi identifikasi judul berita *hoax* berbahasa Indonesia.

## 1.4 Batasan Masalah

Peneliti menetapkan batasan pada proses penelitian untuk menghindari kesalahan dalam penelitian, termasuk masalah berikut ini :

1. Data yang digunakan hanya bahasa Indonesia.
2. Data yang digunakan yaitu data yang bertemakan kesehatan.
3. Data diambil dari rentang tahun 2018 sampai tahun 2022.

4. *Output* pada penelitian yaitu sistem yang berbasis *web* untuk mengidentifikasi judul berita apakah *hoax* atau *non-hoax*.

## 1.5 Manfaat Penelitian

Beberapa manfaat yang diperoleh pada penelitian ini adalah :

1. Membantu masyarakat agar lebih mudah mendapatkan informasi yang benar dan terhindar dari berita-berita *hoax* yang beredar.
2. Menjadi sumber penelitian selanjutnya dalam melakukan identifikasi berita *hoax* berbahasa Indonesia menggunakan *Bidirectional Long Short Term Memory.*

## 1.6 Metodologi Penelitian

Proses pada penelitian meliputi :

1. Studi Literatur

   Dalam tahapan ini, informasi referensi dikumpulkan dari buku, *website*, jurnal, artikel, serta sumber bacaan lainnya yang bertentangan dengan berita *hoax, Natural Language Processing, text processing*, dan metode *Bidirectional Long Short Term Memory.*
2. Analisis Permasalahan

   Pada langkah ini, referensi yang dikumpulkan pada langkah sebelumnya, yaitu studi literatur, dianalisis untuk memperoleh pemahaman tentang *Natural Language Processing* yang akan diterapkan dalam penelitian untuk mengidentifikasi berita *hoax* berbahasa Indonesia menggunakan *Bidirectional Long Short Term Memory* (Bi-LSTM).
3. Perancangan Sistem

   Berdasarkan hasil analisis permasalahan pada langkah sebelumnya, dilakukan perancangan sistem, penentuan pengujian data, serta perancangan arsitektur.
4. Implementasi

Pada tahap ini dilakukan untuk mengimplementasi kode program berdasarkan tahap analisis dan perancangan sistem yang sudah dibuat sebelumnya.

5. Pengujian Sistem

   Selanjutnya, sistem diuji sehingga sistem yang sudah dibuat dapat digunakan dengan metode *Bidirectional Long Short Term Memory* dalam mengidentifikasi berita *hoax*.

6. Penyusunan Laporan

   Pada tahap akhir, dilakukan penyusunan laporan dari hasil analisis keseluruhan yang dilakukan melalui penerapan metode *Bidirectional Long Short Term Memory*.

## 1.7 Sistematika Penulisan

Berdasarkan cara sistematis, dalam penelitian ini terdiri dari 5 bab meliputi :

**BAB 1 PENDAHULUAN**

Dalam bab ini membahas latar belakang penelitian, termasuk rumusan, tujuan, batasan, dan manfaat dari penelitian, serta metodologi dan sistematika penulisan.

**BAB 2 LANDASAN TEORI**

Bab ini menyajikan beberapa teori yang dibutuhkan guna mengetahui subjek penelitian ini.

**BAB 3 ANALISIS DAN PERANCANGAN SISTEM**

Proses analisis dan perancangan sistem, serta arsitektur umum, hendak dijelaskan dalam bab ini.

**BAB 4 IMPLEMENTASI DAN PENGUJIAN**

Bab ini membahas pelaksanaan rancangan yang dijabarkan pada Bab 3 serta pengujian sistem untuk mengidentifikasi keunggulan dan kekurangan sistem.

## BAB 5 KESIMPULAN DAN SARAN

Bab 3 membahas rangkuman rancangan, Bab 4 menjelaskan hasil, dan Bab akhir yaitu Bab 5 memberikan saran untuk penelitian lanjutan serta kesimpulan dari penelitian yang dibuat oleh penulis.

# BAB 2

# LANDASAN TEORI

## 2.1 *Text Mining*

Menurut Hearst (2009) *text mining* adalah cara mengekstraksi data otomatis oleh berbagai sumber teks untuk menghasilkan informasi baru yang belum pernah diketahui komputer sebelumnya. Elemen kuncinya adalah menghubungkan informasi yang diekstraksi bersama-sama untuk membentuk fakta atau hipotesis baru untuk dieksplorasi lebih lanjut dengan cara eksperimen yang lebih konvensional.

Pada saat ini, *text mining* dibutuhkan untuk memvisualisasikan ataupun mengevaluasi informasi yang diperoleh dari gabungan dokumen tulisan yang sangat besar. *Text mining* merupakan proses untuk mendapatkan informasi berkualitas tinggi pada teks, umumnya dengan menggunakan pola statistik untuk memperhatikan pola dan tren. *Text mining* memberikan nilai atau bobot pada *term* dalam dokumen dengan menggunakan bobot kata. Bobot yang diberikan pada *term* tergantung pada metode yang akan digunakan. (Deolika dkk., 2019).

*Text mining* sangat penting untuk pengembangan aplikasi karena memungkinkan untuk mengetahui isi teks secara langsung tanpa membaca teks secara terpisah. *Text mining* adalah proses pengolahan data, tetapi tidak untuk data yang dikelola seperti teks yang tidak tersusun atau setengah tersusun contohnya teks email dan HTML, serta teks yang ditemukan di beberapa sumber (Jaka H, 2015).

Menurut Wijaya & Santoso (2016) terdapat 3 proses yang biasa dilakukan dalam *text mining* yaitu sebagai berikut :

1. *Characterization of data*

   Karena tidak perlu menggunakan banyak tag HTML, teks distrukturkan terlebih dahulu sebelum dimasukkan ke dalam database melalui *parsing.*

2. *Data mining*

   Selanjutnya, pencarian dilakukan menggunakan algoritma tertentu dari data yang ada untuk menghasilkan model dari data tersebut.

3. *Data visualization*

Hasil pencarian yang ada akan menghasilkan teks yang mudah dipahami.

## 2.2 Identifikasi

Menurut KBBI (2018) identifikasi merupakan bentuk penentuan yang dihasilkan dari penetapan identitas seseorang atau benda dalam penanganan masalah sosial tertentu. Identifikasi merupakan penentuan atau penetapan identitas seseorang dan proses mengidentifikasi adalah kegiatan untuk menentukan atau menetapkan identitas seseorang. Pengembangan teknologi identifikasi ini diterapkan pada berbagai perangkat salah satunya pada *smartphone* berbasis *Android* (Widiakumara dkk., 2017). Kartini Kartono (2018) mengemukakan bahwa identifikasi adalah proses sosial dan interaksi sosial yang membuat serangkaian pengenalan untuk menempatkan obyek dalam suatu kelas sesuai dengan karakteristik tertentu. Dari pendapat para ahli tersebut, dapat disimpulkan bahwa pengertian identifikasi adalah cara yang dilakukan oleh individu dalam pengambilan ahli karakteristik seseorang. Identifikasi merupakan tindakan yang dilaksanakan dengan proses mencari, mendapatkan, memeriksa, dan mencatat informasi tentang sesuatu atau seseorang. Proses identifikasi teks sangat penting dan bertujuan untuk mengenali pola teks yang akan diklasifikasikan dan mengenali jenis teks yang akan dipergunakan sebagai *training.*

## 2.3 Berita

Berita merupakan salah satu bentuk informasi yang sering ditemukan dari cara penyebarannya. Berita disajikan dengan gaya dan bahasa tersendiri. Beragamnya gaya dan bahasa bertujuan agar informasi yang disajikan dapat diterima dan menarik oleh seluruh lapisan masyarakat (Retnowati, 2019).

Berita yang kita baca setiap hari di surat kabar, majalah, buletin, dan media visual dan audio visual lainnya adalah hasil dari konflik antara nilai-nilai masyarakat dan aturan yang diterapkan di media (Mahdi, 2015).

Berita mempunyai beberapa ciri yang digunakan dalam teks berita yaitu sebagai berikut :

1. Berdasarkan fakta bukan pendapat dari penulis.
2. Bahasa yang digunakan mudah dipahami dan menarik para pembaca.

3. Data yang disajikan sesuai dengan konteks dan lengkap.
4. Bersifat objektif yaitu sesuai dengan kejadian yang berlangsung.
5. Sumber berita harus benar dan bisa dipertanggungjawabkan.

## 2.4 *Hoax*

*Hoax* berasal dari istilah *(hocus to trick)* yang diciptakan untuk memanipulasi seseorang atau mengajak seseorang untuk melakukan suatu tindakan menggunakan ancaman ataupun penipuan. Motif *hoax* dapat bersifat komersial dan politis, dan dapat menyebabkan dampak buruk seperti hilangnya reputasi, materi, bahkan mengancam nyawa (Prasetijo dkk., 2019).

*Hoax* adalah informasi yang tidak dapat dipercaya karena yang disampaikan adalah informasi palsu tetapi dianggap sebagai kebenaran. *Hoax* mampu mempengaruhi reputasi dan kepercayaan pada banyak orang. Berita *hoax* menyebar lebih cepat daripada berita sebenarnya (Ismayanti & Setiawan, 2021).

*Hoax* dapat memberikan pengaruh buruk pada seseorang melalui tulisan dan dapat mempengaruhi pikiran waras seseorang. Sementara itu, gambar dapat memunculkan rasa takut serta terancam. Lembaga Ilmu Pengetahuan Indonesia (LIPI) menyatakan bahwa berita *hoax* lebih rentan terjadi pada masyarakat yang fanatik (Pardede & Ibrahim, 2020).

Menurut Rahadi (2017) ada beberapa jenis informasi *hoax* yang akan dijelaskan sebagai berikut :

1. *Fake News*

   *Fake news* merupakan berita yang berusaha menggantikan berita yang asli. Berita ini bertujuan untuk memalsukan atau memasukkan ketidakbenaran dalam suatu berita. Penulis berita bohong biasanya menambahkan hal-hal yang tidak benar.

2. *Clickbait*

   *Clickbait* merupakan tautan yang diletakkan secara strategis di dalam suatu situs dengan tujuan untuk menarik orang masuk ke situs lainnya. Konten di dalam tautan ini sesuai fakta namun judulnya dibuat berlebihan atau dipasang gambar yang menarik untuk memancing pembaca.

3. *Confirmation Bias*

   *Confirmation bias* adalah kecendrungan untuk menginterpretasikan kejadian yang baru terjadi sebaik bukti dari kepercayaan yang sudah ada.

4. *Misinformation*

   *Misinformation* adalah informasi yang salah atau tidak akurat, terutama yang ditujukan untuk menipu.

5. *Satire*

   *Satire* berarti sebuah tulisan yang menggunakan humor, ironi, hal yang dibesar-besarkan untuk mengomentari kejadian yang sedang hangat.

6. *Post-truth*

   *Post-truth* yaitu kejadian dimana emosi lebih berperan daripada fakta untuk membentuk opini publik.

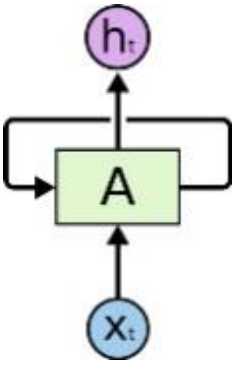
7. *Propaganda*

   *Propaganda* merupakan aktivitas menyebarluaskan informasi, fakta, argumen, gosip, setengah kebenaran, atau bahkan kebohongan untuk mempengaruhi opini publik.

Pada penelitian ini, jenis informasi hoax yang termasuk adalah *Fake News*, dimana penelitian ini mengidentifikasi teks berita palsu yang beredar.

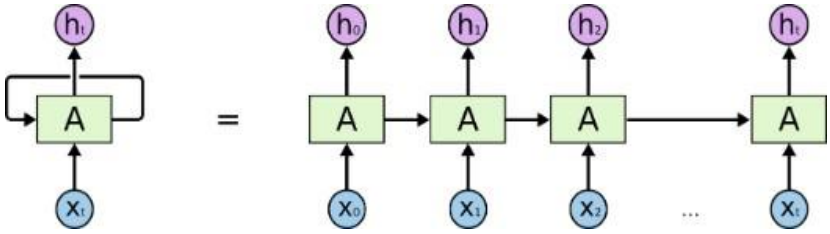
## 2.5 *Reccurent Neural Network* (RNN)

*Recurrent Neural Network* adalah suatu metode yang telah muncul dari tahun 1980-an. *Reccurent Neural Network* adalah sistem yang dirancang khusus yang digunakan untuk memproses data secara berurutan atau yang disebut *sequential data*. RNN juga dapat diartikan sebagai proses yang dapat mengolah *input* dengan beberapa informasi yang sudah ada sebelumnya. RNN mempunyai tugas untuk menyimpan sebuah memori agar dapat mengenali data dengan baik, lalu menggunakannya dengan cara membuat prediksi yang tepat (Yanuar, 2018). Adapun gambaran proses RNN terlihat pada Gambar 2.1.




**Gambar 2.1** *Looping* Informasi dalam Metode RNN (Olah, 2015)

Gambar 2.1 menunjukkan bahwa *Xt* adalah masukan dan *Ht* adalah keluaran, serta jalur pengulangan mengharuskan data ditransfer oleh satu proses jaringan menuju proses selanjutnya. Terdapat beberapa salinan dari jaringan yang sama disebut sebagai *Recurrent Neural Network.* Jaringan-jaringan tersebut akan mengirimkan sebuah pesan kepada jaringan setelahnya sebagaimana ada di Gambar 2.2.



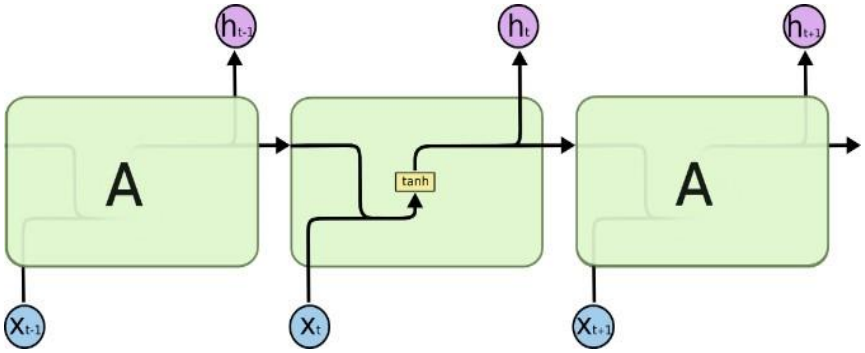

**Gambar 2.2** RNN terdiri dari Berbagai Salinan Jaringan Serupa (Olah, 2015)

Terdapat masalah yang ada di arsitektur *Reccurent Neural Network* yaitu ketergantungan jangka panjang. Untuk mengatasi masalah tersebut digunakan LSTM yang merupakan variasi dari RNN.

### 2.6 *Long Short Term Memory* (LSTM)

*Long Short Term Memory* ialah *reccurent neural network* dan dilengkapi oleh mekanisme gerbang khusus yang mengontrol akses ke sel memori. Pada prinsipnya, *reccurent networks* dapat menggunakan koneksi umpan baliknya untuk menyimpan representasi peristiwa *input* terbaru dalam bentuk aktivasi (*short-term memory*, sebagai lawan dari *long-short memory* yang diwujudkan dengan bobot yang berubah

secara perlahan) (Hochreiter & Schmidhuber, 1997). Pada RNN, *layer tanh* hanya dipakai satu *layer* yang sederhana diperulangan jaringan ditunjukkan melalui Gambar 2.3.




**Gambar 2.3** Lapisan *Tanh* dalam Metode RNN (Olah, 2015)

Persamaan *tanh* dijelaskan dalam Persamaan 2.1.

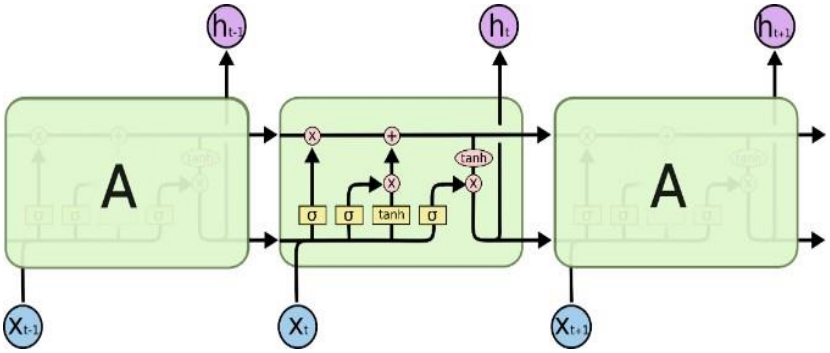
$$\tanh(x) = 2\sigma(2x) - 1 \qquad (2.1)$$

Keterangan :

$\sigma$ = sifat *sigmoid*

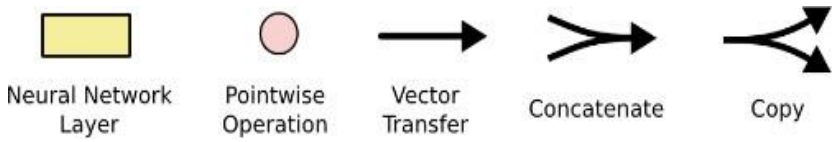
$x$ = inputan data

*Long Short Term Memory* (LSTM) mempunyai model perulangan berupa empat lapisan seperti pada Gambar 2.4.




**Gambar 2.4** *Looping* Empat Lapisan Metode LSTM (Olah, 2015)

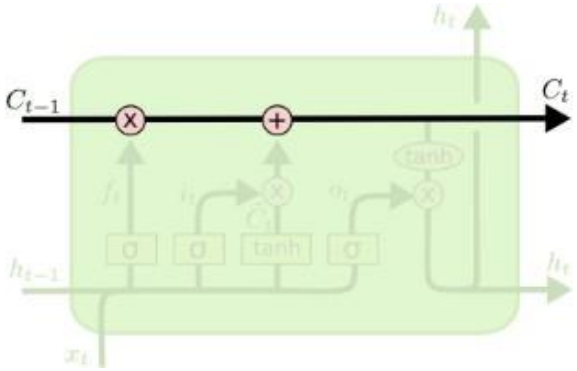
Beberapa notasi yang dapat digunakan dari diagram LSTM seperti ditunjukkan pada Gambar 2.5.




**Gambar 2.5** Notasi dalam Metode LSTM (Olah, 2015)

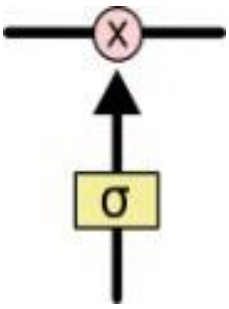
Setiap baris pada diagram membawa semua vektor, dari *output* satu ke *input* lainnya. Kotak kuning merupakan lapisan jaringan saraf, kemudian lingkaran merah muda disebut operasi titik, contohnya penjumlahan vektor. *Concatenate* yaitu penggabungan garis, sedangkan *Copy* menunjukkan percabangan garis yang berarti kontennya akan disalin dan pergi ke lokasi yang berbeda.

Dalam LSTM kunci paling pertama yaitu *cell state* yang memiliki arti sebagai garis horizontal pada LSTM yang digunakan untuk menyatukan seluruh *output layer* melalui Gambar 2.6.




**Gambar 2.6** *Cell State* dalam Metode LSTM (Olah, 2015)

Pada Gambar 2.7, Kemampuan LSTM adalah menghapus atau menambahkan informasi ke c*ell state* yang dikatakan sebagai *gates*. *Gates* mempunyai *pointwise multiplication operation* dan *sigmoid neural net layer.*

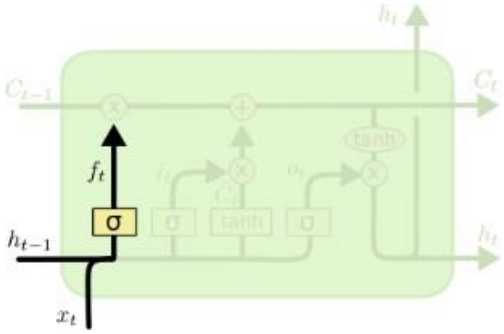



**Gambar 2.7** *Sigmoid Layer* dalam LSTM (Olah, 2015)

Dalam sistem *Long Short Term Memory* (LSTM) terdapat empat tahapan, dimana dari keempat tahap tersebut masing-masing mempunyai tugas dan kegunaannya dalam memproses data, mengklasifikasi, dan mengumpulkan.

1. *Forget Gate*

   *Forget gate* merupakan *gate* pertama dalam LSTM. Tugas dari *forget gate* adalah untuk melupakan beberapa informasi yang tidak diperlukan atau didalam sebuah sistem. Input $h_{t\_1}$ dan $x_t$ akan digunakan pada *forget gate*, dan angka 0 dan 1 akan dihasilkan sebagai *output* di dalam *cell state* $C_{t\_1}$, berikut disajikan melalui Gambar 2.8.




**Gambar 2.8** Tahap Lapisan *Forget Gate* untuk Metode LSTM (Olah, 2015)

Rumus persamaan *forget gate* dapat ditunjukkan dalam Persamaan 2.2.

$$f_t=\sigma\,(wf.[h_{t\_1},x_t]+bf) \tag{2.2}$$

Penjelasan :

$f_t$ = *forget gate*

$\sigma$ = sifat *sigmoid*

*wf* = *weight* pada *forget gate*

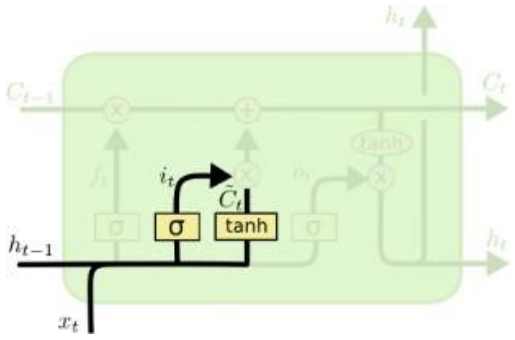
$h_{t\_1}$ = keluaran sebelum orde ke t

$x_t$ = masukan untuk orde ke t

*bf* = bias untuk *forget gate*

2. *Input Gate*

*Input gate* ialah gerbang yang berfungsi mendukung keakuratan data dengan memasukkan informasi yang telah dipilih sebelumnya melalui *forget gate* dan memungkinkan *forget gate* untuk melakukannya. Pada tahap ini, ada dua bagian. Seperti ditunjukkan pada Gambar 2.9, komponen pertama yaitu lapisan *input gate*, yang menentukan nilai yang akan diperbaiki. Selanjutnya, lapisan *tanh* menentukan nilai baru, yaitu $\acute{C}_t$,yang merupakan hasil *output* dari lapisan *input gate*. Kedua lapisan ini dihubungkan untuk memperbaiki *cell state.*




**Gambar 2.9** Tahap *Input Gate Layer* & *Tanh Layer* Metode LSTM (Olah, 2015)

Rumus persamaan *input gate* dapat ditunjukkan dalam Persamaan 2.3.

$$i_t=\sigma(w_i.[h_{t\_1},x_t] + b_i) \quad (2.3)$$

Penjelasan :

$i_t$ = *input gate*

$\sigma$ = sifat *sigmoid*

$w_i$ = *weight* pada *input gate*

$h_{t\_1}$ = keluaran sebelum orde ke t

$x_t$ = masukan untuk orde ke t

$b_i$ = bias untuk *input gate*

Adapun persamaan 2.4 berikut menunjukkan rumus persamaan kandidat.

$$\acute{C}_t = \tanh(wc.[h_{t\_1}, x_t] + bc) \quad (2.4)$$

Penjelasan :

$\acute{C}_t$ = nilai kandidat yang dimasukkan di *cell state*

*tanh* = fungsi dari *tanh*

*wc* = *weight* pada *cell state*

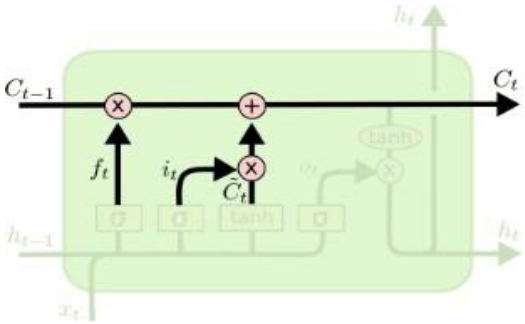
$h_{t\_1}$ = keluaran sebelum orde ke t

$x_t$ = masukan untuk orde ke t

*bc* = bias pada *cell state*

3. *Cell State*

Pada tahap ini, nilai *cell state* lama yaitu $C_{t\_1}$ akan diubah menjadi nilai *cell state* baru, Ct, yakni ditunjukkan melalui Gambar 2.10. Tahap ini digunakan sebagai memori *layer* dan berfungsi untuk mengingat informasi yang berlangsung lama.




**Gambar 2.10** Tahap *Cell State* dalam Metode LSTM (Olah, 2015)

Adapun rumus persamaan *cell state* ditunjukkan dalam Persamaan 2.5.

$$C_t = f_t * C_{t\_1} + i_t . \acute{C}_t \quad (2.5)$$

Penjelasan :

$C_t$ = *cell state*

$f_t$ = *forget gate*

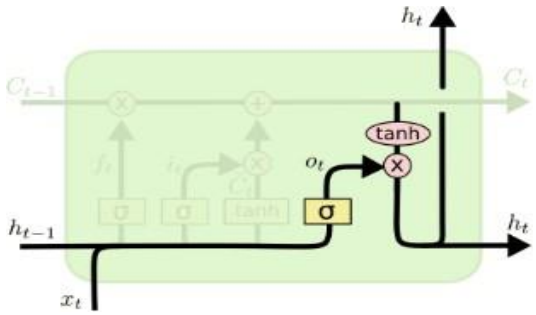
$C_{t\_1}$ = nilai *cell state* sebelum orde ke t

$i_t$ = *input gate*

$\acute{C}_t$ = nilai konteks untuk dimasukkan di *cell state*

4. *Output Gate*

*Ouput gate* merupakan gerbang terakhir pada LSTM yang berguna untuk menentukan apakah telah dibangun pada *input* maupun *cell gate*. *Output gate* menunjukkan pembagian nilai dari *memory cell* pada waktu *t* + 1 dan seterusnya. Tahap *output gate* terdapat dalam Gambar 2.11.




**Gambar 2.11** Tahapan *Output Gate* dalam LSTM (Olah, 2015)

Rumus persamaan *output gate* dapat ditunjukkan dalam Persamaan 2.6.

$$o_t = \sigma(w_o.[h_{t\_1}, x_t] + b_o) \quad (2.6)$$

Penjelasan :

$o_t$ = *ouput gate*

$\sigma$ = sifat *sigmoid*

$w_o$ = *weight* pada *output gate*

$h_{t\_1}$ = keluaran sebelum orde ke t

$x_t$ = masukan untuk orde ke t

$b_o$ = bias pada *output gate*

Persamaan 2.7 di bawah ini menunjukkan rumus persamaan *output* orde t.

$$h_t = o_t * \tanh(c_t) \quad (2.7)$$

Penjelasan :

$h_t$ = *ouput* pada orde ke t

$o_t$ = *output gate*

*tanh* = fungsi dari *tanh*

($c_t$) = *cell state*

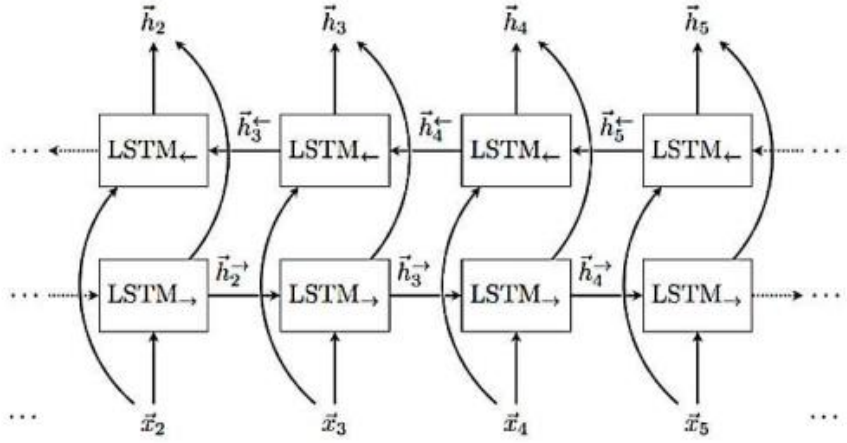
## 2.7 *Bidirectional Long Short Term Memory* (Bi-LSTM)

*Bidirectional Long Short Term Memory* (Bi-LSTM) merupakan jaringan syaraf dari *Long Short Term Memory* (LSTM) yang terdiri dari dua lapisan jaringan syaraf tiruan LSTM, yaitu lapisan LSTM maju yang berfungsi untuk memodelkan konteks sebelumnya serta lapisan LSTM mundur yang berfungsi untuk memodelkan setiap konteks berikutnya.

Menghubungkan dua lapisan tersembunyi ke *output* yang sama dari arah yang berlawanan adalah inti dari *Bidirectional* LSTM (Isnain dkk., 2020). *Output* pada lapisan ini umumnya digabungkan menjadi satu. Dengan lapisan ini, model dapat mempelajari informasi masa lalu (*past*) dan informasi masa mendatang (*future*) untuk tiap sekuen input.

Pada LSTM hanya dapat mengetahui data informasi dari data sebelumnya. Sehingga tidak mengenali informasi yang akan datang. Maka dari itu dihadirkan Bi-LSTM yang dapat bekerja dari dua arah (*Bi-directional*) yang menggabungkan informasi LSTM *forward* dan LSTM *backward*.

Arsitektur dalam Bi-LSTM terdiri dari LSTM *forward* dan LSTM *backward.* Bi-LSTM dapat menyesuaikan data dari arah maju (*forward*) dan mundur (*backward*), lalu menggabungkan prediksi. *Forward* serta *backward* pada Bi-LSTM dapat meningkatkan jumlah informasi yang ada ke jaringan serta konteks yang tersedia untuk algoritma. Berikut arsitektur *Bidirectional* LSTM terdapat pada Gambar 2.12.




**Gambar 2.12** Arsitektur *Bidirectional* LSTM (Lample dkk, 2016)

## 2.8 *Word Embedding*

*Word embedding* merupakan teknik dari *Natural Language Processing* (NLP) yang berfungsi sebagai proses konversi kata-kata seperti karakter *alphanumeric* di dalam sebuah vektor sehingga dapat diproses oleh algoritma *machine learning*. *Word embedding* bertujuan untuk memudahkan proses analisis teks agar lebih efektif dan meningkatkan akurasi pada model *machine learning*. Dengan melakukan teknik *word embedding,* kata-kata yang mempunyai konteks yang serupa akan diletakkan pada wilayah yang bersanding satu sama lain di suatu ruang vektor. Pada penelitian ini teknik *word embedding* yang dipakai adalah *word2vec.*

## 2.9 Penelitian Terdahulu

Penelitian sebelumnya seperti yang dilakukan oleh Al-Ash & Wibowo (2018). Penelitian ini dilakukan untuk mengidentifikasi karakteristik berita palsu. Karakteristik berita palsu direpresentasikan dalam vektor dokumen. Istilah vektor frekuensi dapat digunakan untuk mengkarakterisasi dokumen berita palsu karena kinerja yang dicapai dari representasi mencapai 96,74%.

Selanjutnya, Aziz (2019) melakukan penelitian dengan mengidentifikasi apakah berita kesehatan yang masuk adalah berita *hoax* atau fakta. Nilai akurasi optimum yang dihasilkan sebesar 88%, untuk nilai presisi optimum sebesar 83%, sedangkan untuk nilai *recall* optimum sebesar 100%.

Kemudian, penelitian yang dilakukan oleh Rusli dkk. (2020) yaitu mengimplementasikan dan mengevaluasi algoritma *multilayer perceptron* dalam mengklasifikasikan artikel berita untuk mengidentifikasi berita palsu bersama dengan berbagai konfigurasi seperti model *n-gram* dan metode ekstraksi fitur. Nilai presisi dan *recall* masing-masing mencapai 0,84 dan 0,73, dan skor F1 rata-rata makro 0,82.

Adapun penelitian yang menerapkan metode yang sama yakni *Bidirectional Long Short Term Memory* oleh Isnain dkk. (2020) yaitu mendeteksi *tweet* berbahasa Indonesia sebagai ujaran kebencian atau bukan ujaran kebencian dengan akurasi senilai 94,66%, dengan masing-masing nilai presisi 99,08%, *recall* 93,74% dan *F-measure* 96,29%. Sedangkan untuk tiga layer menggunakan Bi-LSTM menghasilkan akurasi sebesar 96,93%.

Selain itu, penelitian yang dilakukan oleh Hilmawan (2022) ialah membuat model klasifikasi untuk memprediksi sarkasme dengan judul berita berbahasa inggris yang menghasilkan tingkat akurasi validasi 82,55%, precision validasi 82,36%, *recall* validasi 79,53%, dan *f1 score* validasi 80,92%.

Adapun Setiawan & Lestari (2022) melakukan penelitian dimana bertujuan untuk membuat model yang dapat melakukan tugas *stance classification* terbaik dalam konteks bahasa Indonesia. Diharapkan model ini dapat membantu dalam memerangi masalah penyebaran berita palsu, terutama di Indonesia.

**Tabel 2.1** Penelitian Terdahulu

| No | Peneliti, Tahun | Judul | Keterangan |
|---|---|---|---|
| 1 | (Herley Shaori Al-Ash & Wahyu Catur Wibowo, 2018) | Fake News Identification Characteristics Using Named Entity Recognition and Phrase Detection | Penelitian ini bertujuan untuk memodelkan *vector* yang dapat mengakomodasi karakteristik berita palsu sebelum diproses lebih lanjut oleh algoritma menggunakan bahasa Indonesia. Dalam penelitian ini, beritapalsu dan berita asli direpresentasikan menurut model ruang vektor. Kinerja yang dicapai dari representasi mencapai 96,74%. |
| 2 | (Thareq Aziz, 2019) | Identifikasi Hoax Pada Artikel Kesehatan Berbahasa Indonesia Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes | Penelitian ini dilakukan untuk mengidentifikasi artikel kesehatan yang masuk apakah tergolong ke dalam berita *hoax* atau fakta. Nilai akurasi optimum yang dihasilkan untuk klasifikasi artikel kesehatan sebesar 88%, untuk nilai presisi optimum sebesar 83%, sedangkan untuk nilai *recall* optimum sebesar 100%. |

| | | | |
|---|---|---|---|
| 3 | (Rusli et al., 2020) | Identifying Fake News in Indonesian via Supervised Binary Text Classification | Penelitian ini meng-implementasikan dan meng-evaluasi algoritma *multilayer perceptron* dalam meng-klasifikasikan artikel berita untuk mengidentifikasi berita palsu bersama dengan berbagai konfigurasi seperti model *n-gram* dan metode ekstraksi fitur. Masing-masing nilai presisi dan *recall* mencapai 0,84 dan 0,73, dan skor F1 ratarata makro 0,82. |
| 4 | (Isnain et al., 2020) | Bidirectional Long Short Term Memory Method and Word2vec Extraction Approach for Hate Speech Detection | Penelitian ini bertujuan untuk mendeteksi ujaran kebencian atau bukan ujaran kebencian *tweet* berbahasa Indonesia dengan mengunakan metode *Bidirectional Long Short Term Memory* dan metode ekstrasi fitur *Word2Vec* dengan arsitektur *Continuous Bag-OfWord* (CBOW). Nilai akurasi 94,66%, dengan masing-masing nilai presisi 99,08%, *recall* 93,74% dan *F-measure* 96,29%. Untuk Bi-LSTM dengan tiga layer memiliki akurasi 96,93%. |

| | | | |
|---|---|---|---|
| 5 | (Muhammad David Hilmawan, 2022) | Deteksi Sarkasme Pada Judul Berita BerbahasaInggris Menggunakan Algoritme Bidirectional LSTM | Penelitian ini membuat model klasifikasi untuk memprediksi sarkasme pada judul berita berbahasa Inggris. Bi-LSTM lebih unggul dibandingkan LSTM, meskipun sedikit Bi-LSTM mampu menghasilkan akurasi validasi 82,55% dan *f1 score* 80,92%, sedangkan pada LSTM mendapatkan akurasi validasi 81,90% dan *f1 score* 80,47%. |
| 6 | (Esther Irawati Setiawan & Ika Lestari, 2022) | Stance Classification Pada Berita Berbahasa Indonesia Berbasis Bidirectional LSTM | Penelitian ini bertujuan untuk membentuk sebuah model yang mampu melakukan tugas *stance classification* terbaik pada konteks bahasa Indonesia. Model ini diharapkan dapat membantu mengatasi per-masalahan penyebaran berita palsu khususnya di Indonesia. |

Perbedaan penelitian yang dilakukan oleh (Herley Shaori Al-Ash & Wahyu Catur Wibowo, 2018) dengan (Thareq Aziz, 2019) terletak pada metode penelitiannya, dimana penelitian (Herley Shaori Al-Ash & Wahyu Catur Wibowo, 2018) menggunakan metode *Named Entity Recognition* dan *Phrase Detection*. Sedangkan penelitian yang dilakukan oleh (Thareq Aziz, 2019) menggunakan kombinasi metode *K-Nearest Neighbor* dan *Naïve Bayes*, serta penelitian (Thareq Aziz, 2019) hanya mengidentifikasi berita *hoax* pada artikel kesehatan saja.

Selanjutnya perbedaan pada penelitian (Isnain et al., 2020) dengan (Muhammad David Hilmawan, 2022) terletak pada *tweet* atau berita yang dideteksi

dan bahasa yang digunakan, dimana pada penelitian (Isnain et al., 2020) mendeteksi ujaran kebencian *tweet* berbahasa Indonesia, sedangkan pada penelitian (Muhammad David Hilmawan, 2022) mendeteksi sarkasme berita berbahasa Inggris. Dan pada penelitian (Isnain et al., 2020) juga menggunakan teknik *Word2vec Extraction Approach.*

Sedangkan perbedaan keseluruhan dari penelitian terdahulu dengan penelitian ini terletak pada metode penelitian yang digunakan, seperti yang dilakukan oleh (Herley Shaori Al-Ash & Wahyu Catur Wibowo, 2018) peneliti tersebut menggunakan metode *Named Entity Recognition* dan *Phrase Detection*, lalu (Thareq Aziz, 2019) menggunakan kombinasi metode *K-Nearest Neighbor* dan *Naïve Bayes*, dan (Rusli et al., 2020) menggunakan metode *Multilayer Perceptron.* Sedangkan pada penelitian ini, peneliti menggunakan metode *Bidirectional Long Short Term Memory* (Bi-LSTM) dan teknik *Word2Vec*. Perbedaan juga terletak pada berita yang dideteksi atau diindentifikasi serta bahasa yang digunakan, dimana pada penelitian (Muhammad David Hilmawan, 2022) mendeteksi sarkasme berita berbahasa Inggris. Sedangkan pada penelitian ini adalah mengidentifikasi berita *hoax* berbahasa Indonesia.

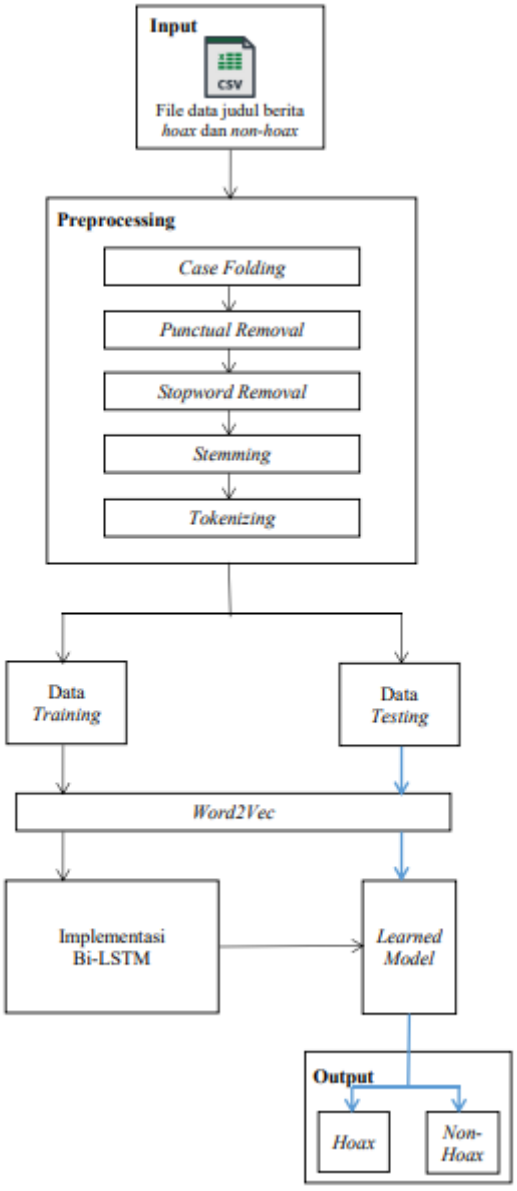
Hasil akurasi yang didapat pada penelitian terdahulu yang sama sama mengidentifikasi berita *hoax* berbahasa Indonesia bertemakan kesehatan seperti yang diteliti oleh (Thareq Aziz, 2019) sebesar 88%, dimana penelitian ini menggunakan kombinasi metode *K-Nearest Neighbor* dan *Naïve Bayes.* Sedangkan pada penelitian ini, akurasi yang berhasil diperoleh sebesar 91% dengan menggunakan metode *Bidirectional Long Short Term Memory* (Bi-LSTM).

# BAB 3

# ANALISIS DAN PERANCANGAN SISTEM

## 3.1 Arsitektur Umum

Arsitektur umum merupakan gambaran secara umum bagaimana cara kerja sistem akan dibangun. Dalam penelitian ini, ada beberapa tahapan yang dilakukan yaitu : dua berkas *file* berekstensi *csv* yang menyimpan data judul berita *hoax* dan *non-hoax* dimuat dalam satu *dataframe*. Selanjutnya *dataframe* tersebut akan masuk ke tahap *preprocessing* yang dimulai dari *case folding, punctual removal, stopword removal, stemming,* dan *tokenizing. Dataframe* dibagi menjadi data *training* dan data *testing*. Kemudian *dataframe* akan diubah formatnya dari text kedalam *word2vec*. Setelah itu model dilatih menggunakan algoritma Bi-LSTM, lalu diakhiri dengan pengujian model untuk mengetahui performa dari model. Adapun arsitektur umum dari rancangan yang digunakan pada penelitian ini dapat dilihat pada Gambar 3.1.




**Gambar 3.1** Arsitektur Umum

### 3.2 Data *Input*

Perancangan sistem pertama kali dimulai dari melakukan data *input*. Data *input* merupakan data yang akan dimasukkan ke dalam sebuah sistem agar dapat diolah menjadi suatu informasi yang berguna. Data input dilakukan agar suatu sistem dapat menghasilkan *output* yang tepat dan bermanfaat untuk pengguna internet. Proses pengolahan data dilakukan sebelum memulai tahap pemodelan atau tahap analisis. Jumlah data yang digunakan sebanyak 8.716 data, kemudian di-*input* ke dalam *file*.csv*. Data judul berita yang termasuk *hoax* bersumber dari portal berita

*Kominfo.go.id* lalu diberi label hoax, sementara data judul berita yang termasuk *non-hoax* bersumber dari *Kemkes.go.id* dan diberi label non_hoax. Data judul berita *hoax* dapat ditunjukkan pada Gambar 3.2, sedangkan data judul berita *non-hoax* ditunjukkan pada Gambar 3.3.

1 Senin 27 April 2020 Sumatera Barat Lockdown
2 Anjuran Berbelanja Ketika Menjalankan "Social Distancing"
3 Kesembuhan Harian Lampaui Penambahan Kasus Baru
4 [Top 5] Topik Teratas Periksa Fakta Mafindo Periode 28 Maret-3 April 2020
5 Beradaptasi Kebiasaan Baru Yang Aman Covid-19 Dan Produktif
6 Lonjakan Kasus Dampak Abaikan Protokol Kesehatan
7 Awas Hati-Hati Apa Yang Disemprotkan Fpi Adalah Virus Corona
8 Kandidat Vaksin Covid-19 Di Indonesia: Cepat, Mandiri Dan Sinergi
9 Mendikbud: Pembukaan Sekolah Tatap Muka Harus Keputusan Bersama
10 Pasien Sembuh Covid-19 Jadi 2.381, Konfirmasi Positif 12.776
11 Universitas Brawijaya Kembalikan Uang Kuliah Akibat Persebaran Virus Corona
12 Pesan Berantai Taman Raya Tahap V Zona Merah Covid-19
13 Perpanjang Psbb Ketat, Elektabilitas Anies Makin Anjlok Gabener Andalan Kadrun Mangkin Ngawur Aja
14 44 Orang Tenaga Kesehatan Positif Covid Di Rs Tipe B Parepare
15 Kandidat Pilkada Harus Contohkan Protokol Kesehatan Yang Ketat

**Gambar 3.2** Data Judul Berita *Hoax*

4359 Waspada Kasus COVID-19 Melonjak, Kemenkes Instruksikan Pemda dan Faskes Siap Siaga
4360 RI-Jepang Perluas Layanan Kardiovaskular di RS Jantung Harapan Kita
4361 Hakordia 2023, Irjen Kemenkes Raih Penghargaan Ahli Pembangun Integritas
4362 Empat RS Vertikal-PT Siemens Healthineers Jalin Kerja Sama Tingkatkan Kompetensi SDM Kesehatan
4363 Antisipasi COVID-19 Jelang Nataru, Kemenkes Tekankan Masyarakat Lengkapi Vaksinasi
4364 Kasus COVID-19 Melonjak Lagi, Yuk Segera Vaksinasi, Gratis!
4365 18 PPPK Kemenkes Dilantik
4366 Kemenkes Raih 3 Penghargaan dalam Ajang TOP DIGITAL Awards 2023
4367 Butuh Kualifikasi Dokter Tinggi, Menkes Minta RSUP dr. Ben Mboi Prioritaskan Dokter Asli NTT
4368 RSUP dr. Ben Mboi Diresmikan Presiden, Warga NTT Tidak Perlu Jauh-Jauh Berobat ke Jakarta
4369 Kasus COVID-19 Naik Lagi, Masyarakat Diminta Tetap Disiplin Prokes
4370 Mycoplasma Pneumoniae Ditemukan di Indonesia
4371 Pabrik Fraksionasi Plasma Pertama di Indonesia Mulai Dibangun
4372 Menkes Kukuhkan Tenaga Cadangan Kesehatan Tipe 2, Targetkan dapat Sertifikasi WHO
4373 Semakin Transparan, Masyarakat Bisa Langsung Cek Stok dan Harga Obat Lewat Farmaplus 2.0
4374 RSUP HAM Kembali Lakukan Operasi Transplantasi Ginjal, Kelima Kali Sejak 2017

**Gambar 3.3** Data Judul Berita *Non-Hoax*

Adapun jumlah masing-masing dari data judul *hoax* dan *non-hoax* disajikan dalam Tabel 3.1.

**Tabel 3.1** Jumlah Dataset *Hoax* dan *Non-Hoax*

| **Dataset** | |
|---|---|
| **Data *Hoax*** | **Data *Non-Hoax*** |
| 4358 | 4358 |

Pada kedua *file* yang telah disusun diatas, selanjutnya akan disatukan menjadi sebuah *dataset* sebelum melakukan proses *prepocessing*. Kemudian pengecekan dilakukan untuk melihat data yang sama. Jika terdapat data yang sama, maka data tersebut akan dihilangkan dan hanya tersisa satu berita.

## 3.3 *Prepocessing*

*Prepocessing* adalah tahap persiapan data sebelum data diproses lebih lanjut. Tahap *prepocessing* dilakukan untuk memeriksa bahwa data atau dokumen yang digunakan dalam analisis sudah bersih dari *noise* dan telah siap untuk digunakan. Tahap *prepocessing* harus dilakukan sebelum pemrosesan data agar mendapatkan hasil akhir yang terbaik dan menghindari kesalahan pada interpretasi data. *Case folding*, *punctual removal*, *stopword removal*, *stemming*, dan *tokenizing* adalah langkah-langkah *preprocessing* dalam penelitian berikut.

### *3.3.1 Case Folding*

Sebuah teks yang diolah menjadi huruf kecil (*lowercase*) akan melalui proses *case folding*, yang memudahkan pencarian dan analisis data. *Case folding* bertujuan untuk menyamaratakan sebuah teks dan memudahkan proses pencarian serta analisis data. Untuk dapat melakukan proses ini digunakan *library* dalam bahasa pemrograman *Python.* Proses *case folding* dilakukan sebelum melakukan analisis teks. Pada penelitian ini, *case folding* digunakan karena adanya ketidak konsistenan penulisan karakter huruf kecil maupun besar pada sebuah data. Tahap *case folding* dapat meningkatkan daya waktu pemrosesan data dikarenakan hanya perlu menyimpan suatu bentuk huruf saja tanpa memperhatikan penulisan huruf tersebut kecil atau besar. Contoh *input* dan *output* proses *case folding* dapat ditunjukkan pada Tabel 3.2.

**Tabel 3.2** Contoh dari *Case Folding*

| **Sebelum Proses *Case Folding*** | **Sesudah Proses *Case Folding*** |
|---|---|
| Daerah yang Panas atau Daerah Bersalju Dapat Membunuh Virus Covid-19 | daerah yang panas atau daerah bersalju dapat membunuh virus covid-19 |

### *3.3.2 Punctual Removal*

*Punctual removal* adalah proses menghilangkan semua tanda baca yang ada di dalam kalimat agar menjadi lebih sederhana. Tanda baca yang dimaksud seperti koma (,), titik (.), tanda seru (!), hastag (#), dan lainnya atau simbol tertentu pada kalimat. *Punctual removal* dilakukan karena terdapat dataset yang masih menggunakan tanda baca atau karakter simbol. Tahap ini bertujuan untuk memproses tanda baca yang tidak berpengaruh serta tidak memiliki arti terhadap teks agar data mudah diolah. *Punctual removal* ini menggunakan *library* pada bahasa pemrograman seperti *Python*. Tahap *punctual removal* dapat dilakukan setelah tahap *case folding*. Contoh *input* dan *output* proses *punctual removal* dapat ditunjukkan pada Tabel 3.3.

**Tabel 3.3** Contoh dari *Punctual Removal*

| **Sebelum Proses *Punctual Removal*** | **Sesudah Proses *Punctual Removal*** |
|---|---|
| daerah yang panas atau daerah bersalju dapat membunuh virus covid-19 | daerah yang panas atau daerah bersalju dapat membunuh virus covid 19 |

### *3.3.3 Stopword Removal*

*Stopword removal* yaitu proses yang dilakukan untuk menghilangkan kata-kata yang kurang penting beserta kata hubung yang tidak bermakna seperti “jika”, “dan”, “andai”,“seperti”,“jadi”, “juga”, “dari”, “yang”, dan lainnya. Tahap ini dilakukan karena *dataset* masih memiliki kata-kata yang tidak penting (*stopword*). *Stopword removal* memiliki tujuan untuk mengurangi kata pembanding dari tiap-tiap kata. Dengan melakukan tahap *stopword removal,* maka hanya kata-kata penting yang menjadi fokus analisis. Tahap ini menggunakan *library* pada bahasa pemrograman *Python* untuk menghapus *stopword* secara otomatis. *Stopword removal* dilakukan setelah tahap *punctual removal. Stopword* yang dihapus dalam penelitian ini dapat ditunjukkan pada Tabel 3.4 dan contoh *input* dan *output* proses *stopword removal* dapat ditunjukkan pada Tabel 3.5.

**Tabel 3.4** Daftar *Stopword* yang Telah Dihapus

| **Daftar *Stopword* yang Telah Dihapus** |
|---|
| ‘yang’, ‘di’, ‘dan’, ‘itu’, ‘dengan’, ‘untuk’, ‘tidak’, ‘ini’, ‘dalam’, ‘akan’, ‘pada’, ‘juga’, ‘karena’, ‘ke’, ‘jika’, ‘menurut’, ‘ia’, ‘para’, ‘sehingga’, ‘ketika’, ‘antara’, ‘namun’, ‘sebagai’, ‘sementara’, ‘kembali’, ‘seperti’, ‘setelah’, ‘bagi’, ‘adalah’, ‘pula’, ‘begitu’, ‘daripada’, ‘terhadap’, ‘kepada’, ‘mengapa’, ‘kenapa’, ‘sebelum’, ‘sesudah’, ‘masih’, ‘kami’, ‘oleh’, ‘saat’, ‘sekitar’, ‘serta’, ‘harus’, ‘hal’, ‘mereka’, ‘dari’, ‘atau’, ‘ada’, ‘telah’, ‘yaitu’, ‘bisa’, ‘bahwa’, ‘sudah’, ‘sambil’, ‘hanya’, ‘maka’, ‘agar’, ‘lagi’, ‘itulah’, ‘kemana’, ‘dimana’, ‘selain’, ‘seolah’, ‘seraya’, ‘supaya’ ‘guna’, ‘secara’, ‘lain’, ‘sedangkan’, ‘seterusnya’, ‘yakni’, ‘melainkan’, ‘sebetulnya’, ‘seharusnya’, ‘anda’, ‘selagi’, ‘toh’, ‘tentang’, ‘tetapi’, ‘walau’, ‘nanti’, ‘supaya’, ‘apakah’, ‘kecuali’, ‘tanpa’, ‘dapat’, ‘setiap’, ‘pun’, ‘kah’, ‘agak’, ‘sebab’, ‘bagaimanapun’, ‘tentu’, ‘amat’, ‘pasti’, ‘saja’, ‘ya’, ‘dsb’, ‘dst’, ‘dll’, ‘demikian’, ‘juga’, ‘mari’, ‘dahulu’, ‘ingin’, ‘tapi’, ‘sesuatu’, ‘setidaknya’ |

**Tabel 3.5** Contoh dari *Stopword Removal*

| **Sebelum Proses *Stopword Removal*** | **Sesudah Proses *Stopword Removal*** |
|---|---|
| daerah yang panas atau daerah bersalju dapat membunuh virus covid 19 | daerah panas daerah bersalju dapat membunuh virus covid 19 |

*3.3.4 Stemming*

*Stemming* yaitu proses menemukan *root* kata oleh setiap kata yang dihasilkan dari *filtering*. Tujuan proses *stemming* adalah untuk menghilangkan imbuhan kata yang ada dalam kalimat, sehingga lebih mudah untuk menemukan kata baru yang akan menjadi kata dasar. Karena *dataset* yang digunakan masih memiliki imbuhan kata,

proses ini dilakukan. Contoh imbuhan kata tersebut seperti "me", "di", "kan", "-pun", "-tah", dan sebagainya. Dalam proses ini digunakan *library* pada bahasa pemrograman yaitu *Python.* Proses *stemming* dilakukan setelah melakukan proses *stopword removal.* Contoh *input* dan *output* proses *stemming* dapat ditunjukkan pada Tabel 3.6.

**Tabel 3.6** Contoh dari *Stemming*

| **Sebelum Proses *Stemming*** | **Sesudah Proses *Stemming*** |
|---|---|
| daerah panas daerah bersalju dapat membunuh virus covid 19 | daerah panas daerah salju dapat bunuh virus covid 19 |

### *3.3.5 Tokenizing*

Untuk membuat analisis data lebih mudah, dibutuhkan proses *tokenizing* guna memecahkan kalimat-kalimat menjadi kata. Proses *tokenizing* bertujuan untuk dapat membedakan yang mana antara pemisah kata atau bukan. Proses ini dilakukan dengan memecahkan deskripsi pada data latih menjadi suatu kata dengan pemotongan *string* pada penyusunnya. Proses *tokenizing* dilakukan setelah melakukan proses *stemming*. Untuk melakukan proses *tokenizing* pada sebuah kalimat digunakan *library* pada bahasa pemrograman *Python*. Contoh *input* dan *output* dari proses *tokenizing* dapat ditunjukkan pada Tabel 3.7.

**Tabel 3.7** Contoh dari *Tokenizing*

| **Sebelum proses *Tokenizing*** | **Sesudah proses *Tokenizing*** |
|---|---|
| daerah panas daerah salju dapatbunuh virus covid 19 | “daerah”, “panas”, “daerah”, “salju”, “dapat”, “bunuh”, “virus”, “covid”, “19” |

## 3.4 *Word2Vec*

*Word2Vec* merupakan model dari *shallow neural network* yang dapat mengubah sebuah representasi kata yaitu kombinasi dari karakter *alphanumeric* ke *vector*. Representasi *vector* mempunyai properti *relationship* pada beberapa kata yang

berhubungan dengan proses *training*. *Word2Vec* adalah metode yang membuat penyisipan kata pada bidang *Natural Language Processing* (NLP). Metode ini diusulkan oleh Tomas Mikolov di Google pada tahun 2013. *Word2Vec* menggunakan berbagai kata dari kumpulan teks untuk masukan yang dapat memberikan representasi *vector*. *Word2Vec* merupakan salah satu teknik *word embedding* yang sering digunakan.

**Tabel 3.8** Contoh Penerapan dari *Word2Vec*

| **Sebelum proses *Word2Vec*** | **Sesudah proses *Word2Vec*** |
|---|---|
| "daerah" | [0.123, 0.456, 0.789] |
| "panas" | [0.987, 0.654, 0.321] |
| "daerah" | [0.123, 0.456, 0.789] |
| "salju" | [0.789, 0.123, 0.456] |
| "dapat" | [0.567, 0.890, 0.123] |
| "bunuh" | [0.456, 0.789, 0.987] |
| "virus" | [0.654, 0.321, 0.987] |
| "covid" | [0.321, 0.789, 0.654] |
| "19" | [0.890, 0.123, 0.567] |

### 3.5 Implementasi Bi-LSTM

Pada penelitian ini, proses yang digunakan untuk implementasi Bi-LSTM adalah sebagai berikut :

1. Membersihkan *Input*

   Proses ini melibatkan pembersihan teks *input* yang diterima dari pengguna. Teks akan melewati fungsi *clean_string()* yang akan menghapus tanda baca, menghapus kata-kata yang tidak relevan atau umum dan mengubah huruf menjadi huruf yang kecil.

2. *Prepocessing* Teks

   Teks yang telah dibersihkan kemudian akan diproses lebih lanjut agar dapat dimasukkan ke dalam model Bi-LSTM. Langkah-langkah ini melibatkan konversi urutan kata menjadi urutan bilangan bulat

menggunakan *tokenizer*, menghitung ukuran *vocab* (jumlah kata yang ada), dan melakukan *padding* pada urutan bilangan bulat agar memiliki panjang yang seragam.

3. Membangun Model Bi-LSTM

   Langkah-langkah ini melibatkan penggunaan *layer embedding* untuk mengubah urutan bilangan bulat menjadi vektor yang saling berkaitan, *layer* Bi-LSTM untuk memproses urutan vektor secara maju dan mundur, dan *layer dense* dengan fungsi aktivasi ReLU untuk memproses *output* dari LSTM.

### 3.6 *Learned Model*

*Learned model* merupakan tahap dalam proses pengerjaan *machine learning*, dimana tahap ini memiliki keunggulan yang baik dalam menginput data dan meningkatkan akurasi yang tergolong tinggi. Dalam tahap ini dilakukan pengujian model pada saat pemrosesan data *training* yang akan dijadikan pembelajaran untuk menghasilkan suatu *output*. Setelah model Bi-LSTM dibangun, teks *input* yang telah dibersihkan melalui formulir *web* akan diproses menggunakan model. Teks *input* akan diubah menjadi urutan bilangan bulat menggunakan *tokenizer* yang sama dengan saat *preprocessing*. Dilakukan *padding* pada urutan bilangan bulat agar memiliki panjang yang seragam. *Input* data tersebut kemudian diberikan ke model Bi-LSTM untuk melakukan prediksi probabilitas (0-1) yang menunjukkan kemungkinan berita tersebut adalah *hoax*.

### 3.7 *Output*

Tahap *output* merupakan tahapan paling akhir dan hasil yang didapatkan adalah identifikasi berita termasuk ke dalam berita *hoax* atau non *hoax* berbahasa Indonesia. Berdasarkan *output* probabilitas dari model, dilakukan klasifikasi dengan mengatur batas probabilitas tertentu. Berita akan dianggap fakta jika probabilitasnya kurang dari 0,5. Jika probabilitas lebih besar dari 0,5, berita akan diklasifikasikan sebagai *hoax*. Hasil klasifikasi ini kemudian ditampilkan pada halaman *web* menggunakan templating *Flask* untuk memberikan informasi kepada pengguna mengenai keaslian berita tersebut.

### 3.8 Perancangan Antar Muka Sistem

Perancangan antarmuka sistem merujuk pada proses merancang tampilan atau pengaturan sistem agar mudah digunakan oleh pengguna. Antar muka sistem berfungsi sebagai titik dimana pengguna berinteraksi dengan sistem melalui perangkat keras serta perangkat lunak. Perancangan antarmuka yang baik dapat memudahkan pengguna dalam mengoperasikan sistem dan meningkatkan efisiensi penggunaan sistem.

1. Rancangan Tampilan

   Pada tampilan ini, perancangan antarmuka sistem identifikasi berita *hoax* berbahasa Indonesia terdiri dari tiga komponen utama yaitu input judul, tombol identifikasi, dan kolom *output* hasil identifikasi seperti terlihat pada Gambar 3.4.


**Gambar 3.4** Rancangan Tampilan

Beberapa penjelasan tentang tiap-tiap komponen tersebut adalah sebagai berikut :

a. *Input* Judul

   Merupakan area di mana pengguna dapat memasukkan judul artikel berita kesehatan yang ingin diperiksa kebenarannya. Pengguna dapat mengetikkan judul tersebut langsung pada kolom *input*, atau melakukan salinan dan tempel (*copy-paste*) dari sumber lain.

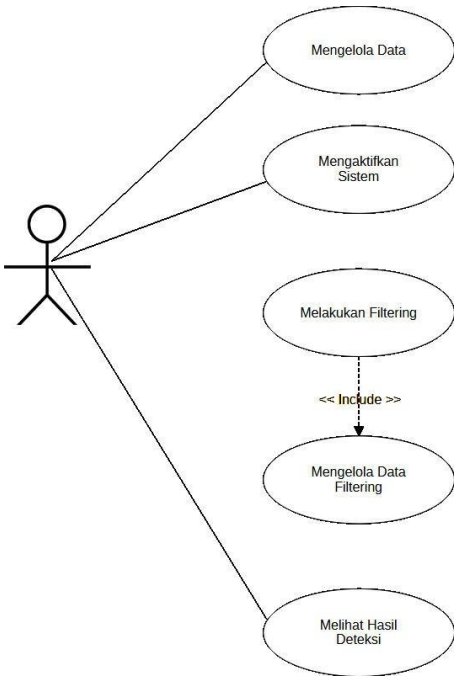
b. Tombol Identifikasi

Setelah judul artikel dimasukkan, pengguna akan menekan tombol identifikasi untuk memulai proses analisis. Tombol ini berfungsi untuk memicu sistem agar memproses data yang dimasukkan oleh pengguna.

c. Kolom *Output* Hasil Identifikasi

Setelah proses analisis selesai, hasil identifikasi akan ditampilkan pada kolom *output*. Pada kolom ini, pengguna dapat melihat apakah artikel berita tersebut dianggap sebagai *hoax* atau tidak oleh sistem.

Adapun *Use Case Diagram* dari sistem tersebut dapat dilihat pada Gambar 3.5 berikut.




**Gambar 3.9** *Use Case Diagram*

Dalam perancangan antarmuka ini, perlu diperhatikan juga faktor keamanan data dan privasi pengguna. Sistem identifikasi berita *hoax* kesehatan ini harus dilindungi dari data dan privasi agar pengguna merasa aman dan nyaman saat menggunakan sistem ini. Karenanya dalam hal ini tidak disertakan opsi untuk memasukkan data pengguna atau mengharuskan pengguna masuk ke sistem dengan identitas atau informasi tertentu.

# BAB 4

# IMPLEMENTASI DAN PENGUJIAN SISTEM

## 4.1 Implementasi Sistem

Setelah membuat antarmuka sistem menggunakan perangkat lunak serta perangkat keras yang akan diuji berdasarkan pada rancangan yang dibahas di Bab 3, tahap selanjutnya akan dilaksanakan.

### *4.1.1 Spesifikasi Perangkat Keras serta Perangkat Lunak*

Perangkat keras berikut ini digunakan untuk menerapkan sistem dalam mengidentifikasi berita *hoax* berbahasa Indonesia :

1. Processor AMD A8 7650K 43 Ghz
2. RAM Team Dark DDR3 8 GB
3. Hardisk 1000 GB

Selain itu, sistem identifikasi berita *hoax* berbahasa Indonesia dibangun menggunakan perangkat lunak seperti :

1. Windows 10 Pro 64 bit
2. Python 3.10
3. Numpy
4. Sklearn
5. Flask
6. Ngrok
7. Jupyter Notebook

### *4.1.2 Implementasi Perancangan Antarmuka*

Implementasi perancangan antarmuka sistem yang sudah dirancang pada Bab 3 sebelumnya akan dijelaskan sebagai berikut :

1. Rancangan Tampilan

   Pada tampilan ini, perancangan antarmuka sistem identifikasi berita *hoax* berbahasa Indonesia terdiri dari tiga komponen utama yaitu *input* judul,

tombol identifikasi, dan kolom *output* hasil identifikasi seperti ditunjukkan pada Gambar 4.1.


**Gambar 4.1** Rancangan Tampilan

Adapun penjelasan untuk masing-masing komponen tersebut adalah sebagai berikut :

a. *Input* Judul

Merupakan area di mana pengguna dapat memasukkan judul artikel berita kesehatan yang ingin diperiksa kebenarannya. Pengguna dapat mengetikkan judul tersebut langsung pada kolom *input*, atau melakukan salinan dan tempel (*copy-paste*) dari sumber lain.

b. Tombol Identifikasi

Setelah judul artikel dimasukkan, pengguna akan menekan tombol identifikasi untuk memulai proses analisis. Tombol ini berfungsi untuk memicu sistem agar memproses data yang dimasukkan oleh pengguna.

c. Kolom *Output* Hasil Identifikasi

Setelah proses analisis selesai, hasil identifikasi akan ditampilkan pada kolom *output*. Pada kolom ini, pengguna dapat melihat apakah artikel berita tersebut dianggap sebagai *hoax* atau tidak oleh sistem.

## 4.2 Pengujian Sistem

Untuk dapat melakukan pemeriksaan pada keseluruhan fungsi sistem yang sudah diimplementasikan, maka perlu dilakukan pengujian sistem. Adapun pengujian sistem identifikasi berita *hoax* berbahasa Indonesia dapat dijabarkan sebagai berikut :

### *4.2.1 Pengujian Input Judul Berita*

Pengujian *input* judul berita dalam sistem identifikasi berita *hoax* berbahasa Indonesia terlihat pada Gambar 4.2.


**Gambar 4.2** Pengujian *Input* Judul Berita

Pada Gambar 4.2 dapat diilihat bahwa *input* judul berita pada sistem dapat dilakukan dengan baik. Sebagai contoh judul berita sebanyak tujuh kata, yaitu “Sendok Bisa Digunakan Untuk Mengecek Kondisi Kesehatan” dapat diketikkan pada sistem tanpa adanya *crash* atau *force close* pada sistem identifikasi berita *hoax* berbahasa Indonesia.

### *4.2.2 Pengujian Deteksi Judul Berita*

Pengujian deteksi judul berita dalam hal ini adalah tahapan pengujian sistem yang paling penting karena merupakan inti dari fungsi sistem yang dirancang, yaitu identifikasi berita *hoax* berbahasa Indonesia, khususnya bertemakan kesehatan. Adapun hasil pengujian deteksi judul berita dapat diuraikan sebagai berikut.

1. Pengujian Pertama Deteksi Judul Berita Dengan *Dataset*


**Gambar 4.3** Pengujian Pertama Deteksi Judul Berita dengan *Dataset* yang Bersumber dari *Kominfo.go.id*

Berdasarkan hasil pengujian pertama deteksi berita *hoax* pada Gambar 4.3 sesuai dengan *dataset* yang bersumber dari *Kominfo.go.id* dengan judul berita "Sendok Bisa Digunakan Untuk Mengecek Kondisi Kesehatan" terlihat bahwa sistem dapat mendeteksi berita *hoax*. Judul berita yang di input sesuai dengan *dataset* yang diperoleh dari *kominfo.go.id* dan memang benar bahwa beritatersebut adalah *hoax.*

2. Pengujian Kedua Deteksi Judul Berita Dengan *Dataset*


**Gambar 4.4** Pengujian Kedua Deteksi Judul Berita dengan *Dataset* yang Bersumber dari *Kominfo.go.id*

Berdasarkan hasil pengujian kedua deteksi berita *hoax* pada Gambar 4.4 sesuai dengan *dataset* yang bersumber dari *Kominfo.go.id* dengan judul

berita "*Website* Mengatasnamakan BPJS Kesehatan" terlihat bahwa sistem dapat mendeteksi berita *hoax*. Judul berita yang di input sesuai dengan *dataset* yang diperoleh dari *Kominfo.go.id* dan memang benar bahwa berita tersebut adalah *hoax.*

3. Pengujian Deteksi Judul Berita Dengan Bahasa Asing


**Gambar 4.5** Pengujian Deteksi Judul Berita Dengan Bahasa Asing

Berdasarkan Gambar 4.5 dapat dilihat bahwa sistem tidak dapat mengidentifikasi judul berita dengan *input* tiga kata bahasa asing, yaitu "*Eat Healthy Fruit*" dengan hasil *input* tidak diketahui, yang artinya sistem berfungsi sesuai dengan *input* bahasa Indonesia atau hanya dapat mendeteksi bahasa Indonesia.

4. Pengujian Pertama Deteksi Judul Berita Dengan Data Diluar *Dataset*


**Gambar 4.6** Pengujian Pertama Deteksi Judul Berita Dengan Data Diluar *Dataset*

Berdasarkan Gambar 4.6 dapat dilihat bahwa pengujian deteksi judul dengan data diluar *dataset*, menunjukkan hasil sistem dapat mengidentifikasi bahwa berita tersebut merupakan fakta atau bukan *hoax*. Judul berita yang diinput merupakan judul berita yang bersumber dari *Suara.com* dan merupakan berita fakta atau bukan *hoax.*

5. Pengujian Kedua Deteksi Judul Berita Dengan Data Diluar *Dataset*


**Gambar 4.7** Pengujian Kedua Deteksi Judul Berita Dengan Data Diluar *Dataset*

Berdasarkan Gambar 4.7 dapat dilihat bahwa pengujian deteksi judul dengan data diluar *dataset*, menunjukkan hasil sistem dapat mengidentifikasi bahwa berita tersebut merupakan fakta atau bukan *hoax.* Judul berita yang diinput merupakan judul berita yang bersumber dari *Kompas.com* yang merupakan berita fakta atau bukan *hoax.*

### 4.2.3 Pengujian Dengan Metode Evaluasi

Pada tahap ini, pengujian dilakukan dengan metode evaluasi atau dikenal sebagai *confusion matrix* sesuai dengan hasil yang diperoleh. Tujuan metode evaluasi ini adalah untuk menjadi alat pengukur dengan melihat bagaimana model bekerja untuk mengetahui apakah judul berita termasuk dalam *hoax* atau tidak. Pada metode evaluasi ini, untuk 8.716 data judul berita, *confusion matrix* diperlukan untuk menghitung tingkat Akurasi, *Precision*, *Recall*, serta *F1 Score*. *Confusion matrix* membantu menggambarkan kinerja model klasifikasi berdasarkan empat hasil klasifikasi: *True*

*Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Pada kasus identifikasi berita ini dapat menggunakan *confusion matrix* dengan label "Prediksi *Hoax*" dan "Prediksi Fakta". Berikut ini dapat ditunjukkan dengan metode evaluasi *confusion matrix* yang terlihat pada Tabel 4.1.

**Tabel 4.1** Keterangan Judul Berita Hoax Dalam *Confusion Matrix*

| **Variabel** | **Jumlah** |
|---|---|
| TP | 3450 |
| TN | 4550 |
| FP | 285 |
| FN | 431 |
| **Total** | **8.716** |

Berdasarkan Tabel 4.1 diatas dapat dilakukan perhitungan Akurasi, *Precision*, *Recall*, serta *F1 Score* dengan menggunakan rumus yang dapat diuraikan sebagai berikut :

Akurasi = $\frac{TP+TN}{TP+TN+FP+FN} = \frac{3450+4550}{3450+4550+285+431} = \frac{8000}{8716} = 0{,}91 \times 100\% = 91\%$

*Precision* = $\frac{TP}{TP+FP} = \frac{3450}{3450+285} = \frac{3450}{3735} = 0{,}92 \times 100\% = 92\%$

*Recall* = $\frac{TP}{TP+FN} = \frac{3450}{3450+431} = \frac{3450}{3881} = 0{,}88 \times 100\% = 88\%$

*F1 Score* = $2 \times \frac{Precision \times Recall}{Precision + Recall} = 2 \times \frac{0{,}92 \times 0{,}88}{0{,}92 + 0{,}88} = 2 \times \frac{0{,}81}{1{,}80} = 0{,}90 \times 100\% = 90\%$

Terdapat beberapa skenario pengujian yang dilakukan dan hasil evaluasi model identifikasi berita *hoax* menggunakan metode *confusion matrix* dengan penggunaan teknik Bi-LSTM. Berikut adalah rincian skenario pengujian yang dilakukan dalam penelitian ini :

1. Pengujian Awal Model

Model dilatih menggunakan *dataset* berita *hoax* dan fakta, kemudian diuji pada *dataset* yang sama untuk melihat kinerja awalnya sebelum optimalisasi.

2. Pengujian dengan *Word2Vec*

   Model diuji dengan menggunakan teknik vektorisasi kata yaitu *Word2Vec*. Teknik tersebut akan diuji untuk membandingkan pengaruhnya terhadap performa model.

3. Pengujian dengan *Dataset* Berbeda

   Model diuji pada *dataset* yang tidak digunakan dalam pelatihan untuk mengukur kemampuannya dalam menggeneralisasi pada data baru.

4. Pengujian dengan Data Pecahan

   Model diuji pada data berita yang lebih sedikit untuk melihat apakah performanya konsisten pada *dataset* yang lebih kecil.

5. Pengujian dengan Data Kontroversial

   Model diuji pada *dataset* berita yang memiliki kontroversi yang lebih tinggi,di mana pemisahan antara *hoax* dan fakta mungkin lebih sulit.

6. Pengujian dengan Data Berita yang Trending

   Model diuji pada berita-berita terkini yang sedang tren, di mana informasi lebih terbatas dan perubahan cepat, sehingga pengujian ini mengukur kecepatan adaptasi model.

Dalam evaluasi kinerja model, *confusion matrix* menggambarkan hasil identifikasi berita oleh model. *True Positive* (TP) adalah jumlah berita yang benar-benar diidentifikasi sebagai *hoax* oleh model, dan mencapai 3450 data. *True Negative* (TN) adalah jumlah berita yang akurat diidentifikasi sebagai fakta, dan berjumlah 4550 data. Namun, terdapat pula *False Positive* (FP), di mana sejumlah 285 berita yang seharusnya adalah fakta, namun disalahidentifikasi sebagai *hoax* oleh model. Sementara itu, *False Negative* (FN) adalah berita-berita yang sebenarnya adalah *hoax,* namun model gagal mengidentifikasinya dan mencapai jumlah 431 data. Dari total 8716 data berita yang diuji, *confusion matrix* ini memberikan

gambaran lengkap mengenai kinerja model dalam mengklasifikasikan berita berdasarkan kebenarannya.

Kemudian dalam menganalisis kinerja model, peneliti juga menggunakan berbagai metrik evaluasi yang memberikan wawasan lebih dalam tentang kemampuan model dalam mengklasifikasikan berita. Pertama, metrik akurasi, yang dalam hal ini menghasilkan angka 91%. Akurasi mengukur sejauh mana model mampu dengan benar mengklasifikasikan baik berita *hoax* maupun fakta. Selanjutnya, presisi atau *precision*, yang dalam hal ini menghasilkan angka 92%. Metrik ini menunjukkan tingkat ketepatan model dalam mengklasifikasikan berita *hoax* dari seluruh berita yang diidentifikasi sebagai *hoax* oleh model. *Recall* atau tingkat kepekaan (*sensitivity*), yang dalam penelitian ini menghasilkan angka 88%. *Recall* mengukur sejauh mana model dapat mengidentifikasi mayoritas berita *hoax* dari total berita yang sebenarnya adalah *hoax*. Terakhir, *F1 Score*,yang memberikan nilai 90%. *F1 Score*, yang menunjukkan tingkat harmonisasi rata-rata antara *Precision* dan *Recall*, memberikan gambaran lengkap tentang seberapa baik model berfungsi dalam mengklasifikasikan berita *hoax* secara keseluruhan.

Analisis hasil metrik evaluasi memberikan wawasan mendalam tentang performa model dalam mengklasifikasikan berita sebagai *hoax* atau fakta. Dari hasil akurasi sebesar 91%, dapat disimpulkan bahwa mayoritas berita berhasil diidentifikasi secara benar oleh model. Tingginya tingkat presisi, mencapai 92%, menunjukkan bahwa ketika model mengidentifikasi suatu berita sebagai *hoax*, peluang kebenarannya cukup tinggi. Meskipun demikian, *recall* yang mencapai 88% menunjukkan bahwa terdapat sejumlah berita *hoax* yang luput dari deteksi model. Namun, *F1 Score* yang mencapai 90% memberikan gambaran komprehensif mengenai performa model, dengan menggabungkan kedua aspek presisi dan *recall*.

Dalam evaluasi kinerja model menggunakan confusion matrix, parameter yang mempengaruhi performa Bi-LSTM meliputi jumlah *neuron*, jumlah *layer*, *dropout rate*, dan panjang urutan *input*. Proses parameter *tuning*, seperti *optimizers*, *learning rate*, *batch size*, dan *epochs*, diperlukan untuk menemukan konfigurasi optimal. Hasil evaluasi model mencakup metrik seperti akurasi, presisi, *recall*, dan F1 *Score*, memberikan pemahaman mendalam tentang kemampuan model mengklasifikasikan berita. Menjelaskan *learned model* melibatkan evaluasi terhadap sejauh mana model dapat memprediksi dengan benar, mengidentifikasi kelemahan atau kekuatan, dan memberikan landasan untuk pengembangan lebih lanjut dalam meningkatkan performa identifikasi berita *hoax*.

# BAB 5

# KESIMPULAN DAN SARAN

## 5.1 Kesimpulan

Hasil penelitian ini memiliki kesimpulan sebagaimana berikut :

1. Penelitian ini mampu mengidentifikasi judul berita *hoax* berbahasa Indonesia secara otomatis.
2. Kinerja metode *Bidirectional Long Short Term Memory*, serta penerapan teknik *Word2Vec* menunjukkan kinerja yang cukup dengan memperoleh tingkat akurasi sebesar 91% dalam mengklasifikasikan berita berdasarkan kategori *hoax* atau fakta yang telah dievaluasi menggunakan metode *confusion matrix.*
3. Sistem dapat melakukan identifikasi terhadap berita fakta yang berasal dari berbagai sumber berita di internet atau situs berita.
4. Berdasarkan pada tingkat akurasi yang diperoleh, maka metode Bi-LSTM, dan teknik *Word2Vec* dapat bekerja dengan baik untuk mengidentifikasi judul berita hoax berbahasa Indonesia.

## 5.2 Saran

Adapula saran dari penulis untuk penelitian seterusnya sebagaimana berikut :

1. Sistem ini hanya menerapkan *dataset* berbahasa Indonesia dan bertemakan kesehatan, sehingga untuk penelitian selanjutnya diharapkan agar dapat dikembangkan untuk dapat mendeteksi berita berbahasa asing dan tema berita yang lebih luas.
2. Pada sistem ini *interface* yang dibuat hanya satu tampilan saja, diharapkan penelitian selanjutnya dapat dikembangkan menjadi beberapa tampilan yang lebih menarik.

# DAFTAR PUSTAKA


Al-Ash, H. S., & Wibowo, W. C. (2018). Fake News Identification Characteristics Using Named Entity Recognition and Phrase Detection. *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 12–17. https://doi.org/10.1109/ICITEED.2018.8534898

Aziz, T. (2019). *Identifikasi Hoax Pada Artikel Kesehatan Berbahasa Indonesia Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes* (Skripsi). Universitas Pembangunan Veteran, Yogyakarta.

Hearst, M. (2009). What Is Text Minning? *SIMS*, *1*(1).

Hilmawan, M. D. (2022). Deteksi Sarkasme Pada Judul Berita Berbahasa Inggris Menggunakan Algoritme Bidirectional LSTM. *Journal of Dinda : Data Science,Information Technology, and Data Analytics*, *2*(1), 46–51. https://doi.org/10.20895/dinda.v2i1.331

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8). Diambil dari https://direct.mit.edu/neco/article-abstract/9/8/1735/6109/Long-Short-Term-Memory? redirectedFrom=fulltext

Ismayanti, F. (2021, Oktober 25). *Deteksi Konten Hoax Berbahasa Indonesia di Twitter Menggunakan Fitur Ekspansi dengan Word2Vec*(Skripsi). Universitas Telkom, S1 Informatika. Diambil dari https://repository.telkomuniversity.ac.id/pustaka/172392/deteksi-konten-hoax-berbahasa-indonesia-di-twitter-menggunakan-fitur-ekspansi-dengan-word2vec.html

Isnain, A. R., Sihabuddin, A., & Suyanto, Y. (2020). Bidirectional Long Short Term Memory Method and Word2vec Extraction Approach for Hate Speech Detection. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems),14*(2), 169–178. https://doi.org/10.22146/ijccs.51743

Kartono, K. (2018). *Patologi Sosial.* Jakarta: PT. Raja Grafindo Persada.

Norton, J. P. (2009). *An Introduction to Identification.* New York: Dover Publications.

Olah, C. 2015. Understanding LSTM Networks. http://colah.github.io/2015-08-Understanding-LSTMs/

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference. https://doi.org/10.18653/v1/n16-1030

Pardede, J., & Ibrahim, R. G. (2020). Implementasi Long Short-Term Memory Untuk Identifikasi Berita Hoax Berbahasa Inggris Pada Media Sosial. *Journal of Computer Science and Informatics Engineering (J-Cosine)*, *4*(2). Diambil darihttps://jcosine.if.unram.ac.id/index.php/jcosine/article/view/361

Prasetijo, A. B., Isnanto, R. R., Eridani, D., Soetrisno, Y. A. A., Arfan, M., & Sofwan, A. (2019). Hoax Detection System on Indonesian News Sites Based on Text Classification using SVM and SGD. *Proc. of 2017 4th Int. Conf. on Information Tech., Computer, and Electrical Engineering (ICITACEE).* Diambil dari http://eprints.undip.ac.id/69088/

Retnowati, J. (2019). *Pengembangan Jenjang Karir Perawat Berbasis Informasi Teknologi terhadap Kinerja Perawat Bagian Kritis di RSUD Dr. Soetomo Surabaya* (Tesis). Universitas Airlangga, Jakarta.

Rusli, A., Young, J. C., & Iswari, N. M. S. (2020). Identifying Fake News in Indonesianvia Supervised Binary Text Classification. *2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and CommunicationsTechnology (IAICT),* 86–90. https://doi.org/10.1109/IAICT50021.2020.9172020

Setiawan, E. I., & Lestari, I. (2022). Stance Classification Pada Berita Berbahasa Indonesia Berbasis Bidirectional LSTM. *Journal of Intelligent Systems and Computation, 1*(1).

Yanuar, A. (2018). Recurrent Neural Network (RNN) – Universitas Gadjah Mada Menara Ilmu Machine Learning. Diambil 21 Februari 2023, dari https://machinelearning.mipa.ugm.ac.id/2018/07/01/recurrent-neural-network-rnn/