SISTEM PENDETEKSI LOKASI SEBARAN PENYAKIT COVID-19 BERDASARKAN INFORMASI KORPUS BERBAHASA INDONESIA DI TWITTER DENGAN PENDEKATAN *EVENT EXTRACTION*

DISERTASI

Untuk Memenuhi Persyaratan Memperoleh Gelar Doktor Ilmu Teknik

FATHONI
03013622126024

Promotor : Prof. Dr. ERWIN., M.Si
Ko-Promotor : Dr. ABDIANSAH.,M.Cs

FAKULTAS TEKNIK
PROGRAM STUDI DOKTOR ILMU TEKNIK
UNIVERSITAS SRIWIJAYA
2024

HALAMAN PENGESAHAN

SISTEM PENDETEKSI LOKASI SEBARAN PENYAKIT COVID-19 BERDASARKAN INFORMASI KORPUS BERBAHASA INDONESIA DI TWITTER DENGAN PENDEKATAN *EVENT EXTRACTION*

DISERTASI

Diajukan untuk memenuhi salah satu syarat memperoleh Gelar Doktor Ilmu Teknik
Pada Fakultas Teknik Universitas Sriwijaya

Oleh

FATHONI
03013622126024

Palembang, 7 Februari 2024

Promotor,	Ko.Promotor,
Prof. Dr. Erwin, S.Si.,M.Si	Dr. Abdiansah, M.Cs
NIP. 197101291994121001	NIP. 198410012009121005

Mengetahui,

Dekan Fakultas Teknik,	**Ketua Program Studi,**
Prof. Dr. Eng. Ir. H. Joni Arliansyah, M.T	Prof. Dr. Ir. Nukman, M.T
NIP. 196706151995121002	NIP. 195903211987031001

HALAMAN PERSETUJUAN

Karya tulis ilmiah berupa Laporan Akhir Disertasi ini dengan judul "Sistem Pendeteksi Lokasi Sebaran Penyakit Covid-19 Berdasarkan Informasi Korpus Berbahasa Indonesia Di Twitter Dengan Pendekatan *Event Extraction*" telah dipertahankan didepan Tim Penguji Karya Tulis Ilmiah Fakultas Teknik Universitas Sriwijaya pada tanggal 7 Februari 2024.

Palembang, 7 Februari 2024
Tim Penguji Karya Tulis Ilmiah berupa Disertasi.

Ketua :

Agung Mataram, S.T, M.T, Ph.D (.....)
NIP. 197901052003121002

Anggota :

1. Prof. Dr. Yusuf Hartono (.....)
   NIP. 196411161990031002
2. Dr. Ermatita. M.Kom (.....)
   NIP. 196709132006042001
3. Dr. Wijang Widhiarso M.Kom. (.....)
   NIDN. 0210097201

Mengetahui,

Dekan Fakultas Teknik,

Prof. Dr. Eng. Ir. H. Joni Arliansyah, M.T
NIP. 196706151995121002

Ketua Program Studi,

Prof. Dr. Ir. Nukman, M.T
NIP. 195903211987031001

SURAT PERNYATAAN

Saya yang bertandatangan di bawah ini:

Nama	:	Fathoni
NIM	:	03013622126024
Program Studi	:	Doktor Ilmu Teknik (Teknik Informatika)
Fakultas	:	Teknik
Perguruan Tinggi	:	Universitas Sriwijaya
Alamat	:	Griya Hero Abadi Blok J No.5 Palembang

Dengan ini menyatakan bahwa Disertasi saya dengan judul "Sistem Pendeteksi Lokasi Sebaran Penyakit Covid-19 Berdasarkan Informasi Korpus Berbahasa Indonesia Di Twitter Dengan Pendekatan *Event Extraction*" merupakan karya sendiri dan bebas dari plagiat. Apabila ditemukan unsur plagiat, maka saya bersedia menerima sanksi akademik sesuai dengan peraturan yang berlaku di Universitas Sriwijaya.

Demikian pernyataan ini dibuat dengan sesungguhnya dan dengan sebenar-benarnya.

Palembang, 07 Februari 2024

Yang Menyatakan,

Fathoni

NIM. 03013622126024

RINGKASAN

Kemunculan Virus Covid-19 pada akhir tahun 2019 di Wuhan China telah membuat kepanikan yang luar biasa di seluruh dunia. Mengacu kepada laporan Aljazera.com tahun 2020 kecepatan penyebaran Virus Covid-19 mencapai 10 kali lipat dibandingkan penyebaran penyakit SARS dan Flu Burung. Kecepatan penularan Virus Covid-19 menyebabkan virus tersebut telah sampai di Indonesia (Jakarta) kurang dari 4 bulan yaitu pada bulan Maret 2020. Banyak strategi yang dapat dilakukan untuk menaggulangi kecepatan penyebaran Covid-19, salah satunya dengan memanfaatkan perkembangan dan pemanfaatan teknologi informasi. Perkembangan ICT ini telah merubah perilaku masyarakat dunia dan Indonesia dalam penyebaran informasi di media sosial online, seperti di Twitter. Berdasarkan pengamatan awal yang peneliti lakukan, banyak informasi cepat yang disampaikan pengguna twitter yang menyampaikan berita penyebaran Covid-19 disuatu lokasi kejadian di Indonesia. Kondisi ini memberikan peluang penelitian baru untuk memanfaatkan dan menangkap informasi lokasi kejadian.penyebaran Covid-19 di Indonesia melalui dataset yang ada di Twitter.

Penelitian ini membahas pengembangan model yang dapat mendeteksi lokasi kejadian peristiwa penyebaran virus covid19 di Indonesia dengan pendekatan *Event Extraction*. Model yang dikembangkan akan melakukan proses ekstraksi kalimat dengan menggunakan metode *Reguler Expression* (Regex) yang telah dimodifikasi dan disesuaikan dengan kebutuhan penelitian dan *Agorithma One-Dimensional (1D) Convolutional Neural Networks* (1D *CNN*).. *Reguler Expression* memproses ekstraksi kalimat data korpus tidak terstruktur yang diperoleh dari server twitter berdasarkan pengetahuan *event trigger* dan *event argument* yang diajarkan kepada model. Untuk menguji performa dan akurasi hasil ekstraksi kalimat yang dilakukan oleh model, peneliti menggunakan metode *Human Intelligence Task* dan *Cross Validation*.

Penelitian ini berhasil mengembangkan model ekstraksi kalimat yang dapat mendeteksi dan mengidentifikasi lokasi kejadian peristiwa penyebaran virus corona di Indonesia dalam format peta digital (spasial) dengan tingkat akurasi 98,58% dan nilai F1-Score 98,92%. Model yang dikembangkan dapat dipergunakan pemerintah dan masyarakat sebagai *early warning* bahwa telah terjadi kejadian peristiwa penyebaran wabah penyakit disuatu lokasi tertentu di Indonesia.

Kata Kunci : Covid-19, *Event Extraction*, Regex, CNN-1D, Indonesia.

SUMMARY

The emergence of the Covid-19 Virus at the end of 2019 in Wuhan, China has created tremendous panic around the world. Referring to the Aljazera.com report in 2020, the speed of the spread of the Covid-19 Virus reached 10 times compared to the spread of SARS and Bird Flu. The speed of transmission of the Covid-19 Virus has caused the virus to have arrived in Indonesia (Jakarta) for less than 4 months, namely in March 2020. Many strategies can be done to respond to the speed of the spread of Covid-19, one of which is by utilizing the development and utilization of information technology. The development of ICT has changed the behavior of the world community and Indonesia in disseminating information on online social media, such as on Twitter. Based on the initial observations that researchers made, a lot of fast information was conveyed by twitter users who conveyed news of the spread of Covid-19 at an incident location in Indonesia. This condition provides new research opportunities to utilize and capture information on the location of the spread of Covid-19 in Indonesia through datasets on Twitter.

This study discusses the development of a model that can detect the location of the occurrence of the spread of the covid19 virus in Indonesia with an Event Extraction approach. The model developed will carry out the sentence extraction process using the Regular Expression (Regex) method that has been modified and adapted to research needs and Agorithma One-Dimensional (1D) Convolutional Neural Networks (1D CNN). Regular Expression processes sentence extraction of unstructured corpus data obtained from twitter servers based on event trigger and event argument knowledge taught to the model. To test the performance and accuracy of sentence extraction results carried out by the model, researchers used the Human Intelligence Task and Cross Validation methods.

This research succeeded in developing a sentence extraction model that can detect and identify the location of the corona virus spread event in Indonesia in digital (spatial) map format with an accuracy rate of 98.58% and an F1-Score value of 98.92%. The model developed can be used by the government and the community as an early warning that there has been an event of disease outbreak spread in a certain location in Indonesia.

Keywords : Covid-19, Event Extraction, Regex, 1D CNN, Indonesia.

BIODATA PENULIS

Fathoni lahir di Palembang, Indonesia, pada tahun 1972. Beliau meraih gelar Sarjana Manajemen Informatika dan Teknik Komputer dari Institut Sains dan Teknologi Akprind Yogyakarta pada tahun 1997 dan Magister Sistem Informasi dari Universitas Gunadarma Jakarta pada tahun 2001. Mulai tahun 1998 dia pernah bekerja sebagai Dosen di program studi Sistem Informasi Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) dan Universitas Bina Darma Palembang, dan pada tahun 2008, ia bergabung sebagai dosen di Jurusan Sistem Informasi Fakultas Ilmu Komputer Universitas Sriwijaya di mana bekerja sampai sekarang. Minat penelitiannya saat ini meliputi bidang *enterprise resource planning* (ERP), *project manajement, dan Data Scince*. Pada tahun 2021, dalam usia yang tergolong tidak muda lagi, dia memutuskan untuk melanjutkan studi pada program doktoral Ilmu Teknik di Fakultas Teknik Universitas Sriwijaya dengan bidang kajian Teknik Informatika. Dia dapat dihubungi di email: fathoni@unsri.ac.id.

KATA PENGANTAR

Segala puji Syukur penulis haturkan kepada Allah SWT yang telah melimpahkan Rahmat, taufiq dan hidayah-Nya serta pertolongan-Nya. Sehingga Disertasi yang berjudul **Sistem Pendeteksi Lokasi Sebaran Penyakit Covid-19 Berdasarkan Informasi Korpus Berbahasa Indonesia Di Twitter Dengan Pendekatan Event Extraction** ini dapat terselesaikan. Shalawat serta salam senantiasa tercurahkan kepada Nabi Muhammad SAW, keluarganya serta sahabatnya yang dinantikan syafaatnya di yaumil akhir.

Dengan tersusunnya Penelitian Disertasi ini, penulis mengucapkan terima kasih dan penghargaan yang setinggi-tingginya kepada Prof. Dr. Erwin, S.Si., M.Si selaku Promotor dan Dr. Abdiansah, M.Cs selaku ko promotor yang telah berkenan memberi bimbingan, arahan dan masukan atas tersusunnya disertasi ini. Penulis juga mengucapkan terima kasih dan penghargaan kepada:

1. Prof. Dr. Taufiq Marwa, SE. M.Si. selaku Rektor Universitas Sriwijaya dan Prof. Dr. Ir. H. Anis Saggaff, M.SCE., MKU., IPU., ASEAN.Eng. APEC.Eng atas kebijakan yang telah ditetapkan, sehingga penulis dapat melanjutkan studi kejenjang strata tiga.
2. Prof. Dr.Eng. Ir. Joni Arliansyah, MT selaku Dekan dan Prof. Dr. Ir. Nukman, M.T selaku Ketua Program Studi Doktor Ilmu Teknik Universitas Sriwijaya atas dukungan dan kebijakan yang ditetapkan sehingga penulis dapat menyelesaikan studi ini.
3. Prof. Dr. Erwin, S.Si., M.Si selaku Dekan dan Alm Dr. Jaidan Jauhari, M.T selaku Dekan periode sebelumnya, atas motivasi dan dukungan kepada saya untuk menempuh studi jenjang doktoral.
4. Ibunda dan Istri tersayang serta anak-anak yang saya cintai yang telah memenuhi hari-harinya dengan Do'a dan harapan semoga penulis dapat menyelesaikan studi ini tanpa hambatan dan rintangan yang berat.
5. Rekan-rekan sejawat di Fakultas Ilmu Komputer Universitas Sriwijaya yang telah banyak memberikan motovasi kepada penulis untuk melanjutkan studi S3.
6. Ibu Yuni selaku staf administrasi prodi S3 fakultas Teknik dan Semua pihak yang telah membantu secara langsung maupun secara tidak langsung yang tidak bisa penulis jelaskan satu persatu.

Penulis menyadari bahwa masih banyak kekurangan dalam penulisan Disertasi ini. Penulis berharap semoga Disertasi ini dapat bermanfaat bagi semua pihak khususnya pada Program Studi Doktor Ilmu Teknik Universitas Sriwijaya.

Salam.

Palembang, 07 Februari 2024.

Fathoni.

DAFTAR ISI

DAFTAR TABEL

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR PERSAMAAN

Halaman

BAB I
PENDAHULUAN

1.1. Latar Belakang.

Kemunculan penyakit baru yang diberi nama *Coronavirus Disease* (COVID19) yang disebabkan oleh virus *Severe Acute Respiratory Syndrome Coronavirus-2* (SARS-CoV-2) [1],[2] yang berasal dari Wuhan China [3] pada pertengahan bulan Desember 2019 telah menghebohkan dunia. Penyebaran penyakit covid19 di Indonesia pertama kali dilaporkan pada awal bulan Maret 2020. Dalam kurung waktu satu bulan, penyakit ini telah menyebar dengan sangat cepat ke berbagai daerah serta menular ke 1.528 orang penduduk Indonesia dengan kasus kematian 136 orang.

Prosedur untuk mendapatkan data masyarakat yang terdampak penyakit covid19 dilakukan melalui mekanisme pendataan secara manual melalui puskesmas, rumah sakit pemerintah dan swasta, klinik kesehatan yang tersebar di Indonesia serta *rapid test* yang dilakukan pada waktu dan lokasi tertentu. Sistem pengawasan seperti ini memerlukan waktu, tenaga kesehatan yang banyak serta biaya yang mahal. Disamping itu, kondisi geografis negara Indonesia yang terdiri dari banyak pulau besar dan kecil serta luasnya wilayah Indonesia memerlukan suatu strategi lain untuk mengetahui dan mempermudah melengkapi data masyarakat yang terdampak covid19, seperti pemanfaatan teknologi informasi.

Perkembangan dan pemanfaatan teknologi informasi yang paling banyak dipergunakan oleh masyarakat umum adalah penggunaan informasi dari berbagai dokumen yang terdapat di suatu *Website*. Kondisi ini menyebabkan peningkatan yang luar biasa terhadap akses pencarian informasi direpositori *web* [4]. Kumpulan dokumen yang semakin lama semakin bertambah volumennya dan secara umum ditampilkankan dalam format yang tidak terstruktur tetap saja menarik perhatian masyarakat dan para peneliti untuk membentuk basis data baru yang berisikan informasi yang dibutuhkan, termasuk juga informasi tentang lokasi penyebaran penyakit Covid19 (Tabel.1.1). Pencarian informasi lokasi suatu kejadian peristiwa di *Website* sangat dimungkinkan karena banyak *web pages* yang membuat berita yang berisikan informasi yang menyebutkan nama tempat kejadian. Hal ini diperkuat dengan hasil pengamatan [5]

yang menyatakan bahwa terdapat 20% dari berita yang dimuat di *web* pernah menyampaikan informasi geografis yang memuat nama tempat kejadian yang dapat dengan mudah dikenali oleh pembacanya. Berita yang menyampaikan informasi tersebut dapat bersumber dari *web* media sosial (seperti : Twitter dan facebook), berita *online* serta situs pemerintahan, dan lain-lain, seperti yang ditampilkan pada Tabel 1.1.

Tabel.1.1. Contoh berita dari *website* (*Twitter*) yang memuat informasi Lokasi Kejadian.

No	Isi Informasi yang disampaikan
1	Tingkat Kesembuhan Pasien Covid19 di Bengkalis Terus Membaik
2	Pasien Positif Covid19 di Kapuas Tersisa 111 Orang
3	557 Pasien Covid19 Sembuh di Aceh, Empat Ribu Dirawat
4	Makin Membaik, Covid19 di Sragen Hanya Tambah 10 Kasus dan Pasien di ICU Tinggal Satu Orang Hari Ini
5	Kasus Aktif Covid19 di Kota Bandung Totalnya 364, Pasien Sembuh Tembus 40 Ribu Orang
6	155 Pasien Covid19 Diisolasi di 5 Fasilitas Isoter di Pekanbaru
7	Empat Ribu Pasien Covid19 di Aceh Masih Dirawat
8	Pasien terkonfirmasi positif COVID19 asal Pamekasan, Pulau Madura, Jawa Timur, yang sedang dirawat di Rumah Sakit Lapangan Indrapura (RSLI) Surabaya melahirkan bayi dengan persalinan normal.
9	Update Data Pasien COVID19 di Kota Medan
10	JUMLAH pasien positif Covid19 di ruangan isolasi RSUD TC.Hillers Maumere terus menurun seiring dengan menurunnya kasus Covid19 di Kabupaten Sikka, Nusa Tenggara Timur. Saat ini, rumah sakit tersebut hanya merawat dua orang pasien Covid19
11	Alhamdulillah, Jumlah Pasien Covid19 yang Dirawat di RS Garut Tinggal 7 Orang

Sumber : Data set Hasil Crawling di website Twitter 2020.

Banyaknya Informasi yang menyampaikan berita yang memuat nama lokasi suatu kejadian di *web* dan media sosial mengindikasikan bahwa kebutuhan pengguna akan informasi geografis sangat besar dan dapat memberikan manfaat dalam kehidupan sehari-hari. Peningkatan kebutuhan informasi ini merupakan peluang baru untuk dilakukan berbagai penelitian terkini tentang teknik atau cara baru untuk melakukan pencarian dan penentuan lokasi kejadian. Peluang ini semakin besar dan menarik karena perkembangan dan penambahan data serta informasi yang tidak terstruktur di *web* bertambah dengan laju yang lebih besar dibandingkan dengan data atau informasi

yang terstruktur. Penelitian terbaru untuk menentukan lokasi kejadian yang bersumber dari data yang tidak terstruktur yang berisikan nama lokasi harus dapat memproses informasi yang terkandung didalam dokumen berita tersebut dan menghasilkan informasi geografis dalam bentuk koordinat atau polygon yang lebih baik dan akurat [6],[7],[8],[9], namun data set yang diolah oleh para peneliti tersebut bersumber dari informasi berita yang sudah dapat dipastikan kebenarannya (berita *online*).

Peningkatan kebutuhan informasi lokasi kejadian suatu peristiwa menyebabkan peluang penelitian dibidang pemrosesan data teks tidak terstruktur dari sumber data selain berita *online*, seperti dari *web* media sosial (seperti Twitter, facebook dan Instagram). Berbeda dengan informasi berita online yang sudah melalui tahap redaksional untuk memastikan kebenaran berita yang akan ditampilkan disuatu *web*, informasi yang disampaikan dimedia sosial masih perlu dipertanyakan kebenaran berita tersebut. Hal ini menarik dan merupakan tawaran penelitian terbaru untuk dilakukan, sehingga informasi lokasi kejadian dari suatu peristiwa yang disampaikan dimedia sosial dapat dipercaya kebenarannya.

Perkembangan dan pemanfaatan teknologi Informasi khususnya *web* media sosial yang terintegrasi dengan teknologi mobile yang sangat pesat didunia termasuk di Indonesia menyebabkan masyarakat sangat mudah untuk mendapatkan serta berbagi informasi secara cepat dan luas. Salah satu tren perkembangan tersebut adalah peningkatan penggunaan *website* dan teknologi mobile yang digunakan masyarakat Indonesia untuk menyampaikan serta berbagi informasi adalah media sosial Twitter. Penduduk Indonesia yang rutin melakukan Twits berjumlah 78 juta orang dari 150 juta pengguna aktif media sosial.

Penggunaan informasi dari Twitter sebagai data set untuk mendeteksi penyebaran penyakit disuatu wilayah atau negara telah banyak dilakukan oleh para peneliti didunia [10]-[13]. Untuk mendapatkan data set yang dibutuhkan di Twitter, para peneliti tersebut melakukan penambangan data dengan cara memanfaatkan fungsi yang telah disediakan oleh Twitter serta mengembangkan sendiri fungsi-fungsi tambahan baru yang disesuaikan dengan karakteristik dan topik penelitian masing-masing. Berdasarkan hasil penelitian yang dipublikasikan, penambangan data set di Twitter yang dikembangkan para peneliti tersebut memiliki tingkat akurasi yang masih

dapat ditingkatkan, terutama untuk hasil penambangan data Twitter penyebaran penyakit covid19 di Indonesia. Kompleksitas untuk mencapai tingkat akurasi penambangan data di Twitter yang baik dipengaruhi oleh beberapa faktor [14][15], antara lain; 1. Kondisi psikologis seseorang (seperti marah dan senang); 2. Karakteristik dan perilaku dari para pengguna Twitter; 3. Sinonim dan Kosakata yang banyak dan bermakna sama; 4. Jumlah total Twit yang dapat diunduh terbatas. Faktor-faktor ini menyebabkan data hasil tangkapan di Twitter mengandung bias dan dapat menyebabkan kesalahan proses pengolahan data lebih lanjut. Permasalahan yang sama terjadi pada proses penambangan data di Twitter untuk mendapatkan data titik lokasi penyebaran penyakit yang disampaikan oleh pengguna Twitter.

Informasi yang bias sampai dengan berita yang tidak benar (*hoax*s) perihal lokasi kejadian peristiwa masih sering disampaikan oleh pengguna Twitter (Tabel 1.2). Informasi yang salah ini tentu saja sangat berpotensi merugikan dan membahayakan para pengguna informasi termasuk juga unsur pemerintahan [16],[17], namun disisi yang lain kondisi ini memberikan peluang penelitian baru untuk menemukan suatu cara atau metode baru yang dapat membantu mengidentifikasi kalimat berita yang mengandung *hoax* tersebut.

Penelitian untuk mengetahui dan mendapatkan lokasi kejadian suatu peristiwa dalam berbagai informasi dalam suatu kalimat telah banyak dilakukan, namun penelitian tersebut masih terfokus kepada penemuan kata yang mengandung makna dari nama lokasi kejadian saja [18][19]. Sedangkan penelitian yang membahas tentang menentukan lokasi kejadian berdasarkan informasi peristiwa kejadian belum banyak dilakukan [20][7][21] dan hal ini memberikan peluang penelitian terbaru untuk menawarkan model dan metode yang cukup atau lebih baik untuk menemukan suatu lokasi kejadian berdasar informasi kalimat yang dituliskan di media *online*. Permasalahan yang paling mendasar dan menantang untuk dilakukan penelitian penentuan lokasi kejadian peristiwa penyebaran penyakit Covid19 adalah tingkat kerumitan yang sangat tinggi yang merupakan interaksi dan pencampuran berbagai aspek seperti ; struktur kalimat Twit yang tidak terstruktur, karakteristik dan keunikan kalimat informasi di Twitter serta aspek temporal dan spasial.

Tabel.1.2. Contoh berita dari *website* (*Twitter*) yang memuat informasi Bias dan Hoax.

No	Isi Informasi yang disampaikan
1	Penurunan Bed Occupancy Rate (BOR) atau ketersediaan tempat tidur untuk pasien Covid19 di Bandar Lampung
2	Manggarai Barat Catat Penurunan Kasus Covid19, Tak Ada Pasien Dirawat di RS
3	Sejak Pandemi terjadi selama 2 thn banyak Oknum RS di Lamongan Bermain tidak benar seperti Memberi obat ke pasien banyak, tidak ada peduli nyawa manusia, pasien non covid malah di covidkan, memaksa orang harus vaksin biar barang cepat habis dan tidak dipikirin nyawa yg penting uang

Kompleksitas permasalahan akan bertambah rumit dengan adanya faktor kebenaran ataupun informasi yang salah perihal telah terjadinya kejadian penyebaran covid yang di-Twit masyarakat di Twitter. Hal ini memberikan peluang untuk menemukan suatu model baru yang dapat mengekstraksi informasi dikalimat *Twit* menjadi informasi yang dapat divalidasi kebenarannya serta mendeteksi lokasi kejadian peristiwa penyebaran penyakit (seperti Covid19) yang dibangun berdasarkan data set ataupun korpus berbahasa Indonesia. Identifikasi lokasi penyebaran penyakit covid secara cepat sangat penting untuk dapat membantu pemerintah Indonesia dalam mendeteksi awal, mencegah penyebaran serta untuk melakukan berbagai tindakan strategis penanggulangan pandemi Covid19 di Indonesia. Hal ini mengisyaratkan korpora berbahasa Indonesia yang akan dibangun dan digunakan untuk melatih data set dalam penelitian ini akan terdiri dari ; 1) Korpus utama yang bersumber dari data set Twit pengguna Twitter di Indonesia dan 2) Korpus *vocabulary* informasi kalimat yang telah tervalidasi dan mengandung makna kata lokasi dan peristiwa kejadian penyebaran penyakit covid19 serta nama-nama kabupaten dan kota yang ada di Indonesia beserta koordinat ***Latitude*** dan ***Longitude*** yang berjumlah 496 kabupaten/kota

Penelitian (disertasi) ini dilakukan untuk mengembangkan model atau metode baru yang dapat mengekstraksi dan mengidentifkasi informasi kalimat kejadian peristiwa serta menangkap data sebaran dan lokasi penyebaran penyakit covid19 di Indonesia dengan tingkat akurasi yang lebih baik dari peneliti sebelumnya. Penelitian penggunaan *Machine learning* untuk membangun model yang dapat mengklasisfikasi dan mengidentifikasi Twit yang mengindikasikan penyebaran penyakit di Indonesia

telah di-lakukan oleh beberapa peneliti sebelumnya [22],[23],[10]. Penelitian yang telah dipublikasikan tersebut menghasilkan klasifikasi Twit adanya penyakit yang cukup baik, namun belum menyampaikan informasi jenis dan lokasi terjadinya penyebaran penyakit, serta tingkat akurasi yang masih dapat ditingkatkan. Pemanfaatan *machine learning* pada penelitian ini juga dimaksudkan untuk membangun model klasifikasi dan pengelompokkan hasil identifikasi Twit data penyebaran penyakit covid19 di Indonesia dengan menggunakan metode yang memiliki tingkat akurasi yang lebih baik dari penelitian sebelumnya. Untuk mencapai tingkat akurasi yang lebih baik, maka penelitian ini menggunakan pendekatan *event extraction* serta melakukan modifikasi algorithma *Reguler Expression* dan menggunakan algorithma *machine learning* yang sesuai dengan karakteristik dan jenis korpus yang telah diperoleh dari *server* Twitter.

1.2. Rumusan Masalah.

1. Bagaimana membangun model ekstraksi kalimat informasi berbahasa Indonesia yang mengandung makna kata lokasi kejadian peristiwa penyebaran penyakit Covid19 di Indonesia berdasarkan data set media sosial Twitter ?
2. Bagaimana membangun peta digital yang dapat mengidentifikasi makna kata lokasi kejadian peristiwa penyebaran covid19 di Indonesia berdasarkan Twit berbahasa Indonesia ?

1.3. Tujuan Penelitian.

1. Mengembangkan model baru yang dapat mengekstraksi kalimat informasi berbahasa Indonesia yang bermakna lokasi kejadian peristiwa penyebaran penyakit Covid19 di Indonesia.
2. Mengembangkan suatu model peta digital yang dapat mengidentifikasi lokasi kejadian peristiwa penyebaran covid19 di Indonesia.

1.4. Manfaat Penelitian.

1. Menghasilkan model baru untuk meningkatkan kualitas ekstraksi kalimat informasi berbahasa Indonesia yang bermakna lokasi kejadian peristiwa

penyebaran virus covid19 dalam kalimat berbahasa Indonesia.

2. Menghasilkan model peta digital yang dapat meningkatkan akurasi identifikasi pendeteksian lokasi kejadian peristiwa dalam kalimat berbahasa Indonesia.
3. Dapat dipergunakan untuk kasus atau kejadian yang lain yang memerlukan deteksi lokasi dari suatu informasi di media sosial.

1.5. Kontribusi Penelitian.

1. Sebagai alternatif solusi untuk mempercepat pemetaan lokasi kejadian terjadinya penyebaran penyakit menular (Covid19) di Indonesia.
2. Menambah kasanah repository korpus berbahasa Indonesia, khususnya untuk mendeteksi lokasi kejadian peristiwa (penyebaran penyakit Covid19).
3. Membantu pemerintah, dalam hal ini Kementerian dan Dinas kesehatan dalam pengambilan keputusan strategis untuk mencegah dan menanggulangi kejadian penyebaran penyakit menular di Indonesia.
4. Membantu masyarakat Indonesia untuk mengetahui lebih cepat dan menghindari lokasi kejadian terjadinya penyebaran penyakit menular (Covid19) di Indonesia.

1.6. Batasan Penelitian

Sumber data korpus utama pada penelitian ini akan diperoleh dari data Twits berbahasa alami Indonesia dari para pengguna Twitter yang ada di server utama Twitter yang berisikan teks berita informasi lokasi dan sebaran penyakit covid19 di Indonesia yang dikumpulkan pada periode bulan Mei, Juni dan Juli tahun 2020. Sedangkan data korpus pelengkap (korpus vocabulary) adalah informasi kalimat yang telah tervalidasi dan mengandung makna kata lokasi dan peristiwa kejadian penyebaran penyakit covid19 serta nama-nama kabupaten dan kota yang ada di Indonesia beserta koordinat ***Latitude*** dan ***Longitude*** yang berjumlah 496 kabupaten/kota. Penelitian ini juga hanya akan melakukan proses pengenalan peristiwa (*event extraction*) terhadap kalimat yang memiliki formula lengkap yang terdiri dari *event trigger* dan *event argument* serta akan tetap memproses dan mengabaikan kosa kata, jumlah kata serta perilaku masyarakat dalam penulisan kalimat Twit di Twitter.

DAFTAR PUSTAKA

[1] B. Ganesh *et al.*, "Epidemiology and pathobiology of SARS-CoV-2 (COVID19) in comparison with SARS, MERS: An updated overview of current knowledge and future perspectives," *Clinical Epidemiology and Global Health*, vol. 10. 2021, doi: 10.1016/j.cegh.2020.100694.

[2] S. Das and A. K. Kolya, "Predicting the pandemic: sentiment evaluation and predictive analysis from large-scale Twits on Covid19 by deep convolutional neural network," *Evol. Intell.*, Mar. 2021, doi: 10.1007/s12065-021-00598-7.

[3] H. A. Rothan and S. N. Byraredd, "Rothan, H. A., & Byrareddy, S. N. (2020). The epidemiology and pathogenesis of coronavirus disease (COVID19) outbreak. Journal of autoimmunity, 102433.," *J. Autoimmun.*, 2020.

[4] H. A. Sleiman and R. Corchuelo, "A class of neural-network-based transducers for web information extraction," *Neurocomputing*, vol. 135, 2014, doi: 10.1016/j.neucom.2013.05.057.

[5] M. Himmelstein, "Local search: The internet is the yellow pages," *Computer (Long. Beach. Calif).*, vol. 38, no. 2, pp. 26–34, 2005, doi: 10.1109/MC.2005.65.

[6] M. B. Imani, S. Chandra, S. Ma, L. Khan, and B. Thuraisingham, "Focus location extraction from political news reports with bias correction," in *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, 2017, vol. 2018-Janua, pp. 1956–1964, doi: 10.1109/BigData.2017.8258141.

[7] M. Karimzadeh, S. Pezanowski, A. M. MacEachren, and J. O. Wallgrün, "GeoTxt: A scalable geoparsing system for unstructured text geolocation," *Trans. GIS*, vol. 23, no. 1, 2019, doi: 10.1111/tgis.12510.

[8] M. Gritta, "Where are you talking about? Advances and Challenges of Geographic Analysis of Text with Application to Disease Monitoring," no. February, p. 140, 2019.

[9] A. Halterman, "Geolocating Political Events in Text," 2019, doi: 10.18653/v1/w19-2104.

[10] M. Adriani, F. Azzahro, and A. N. Hidayanto, "Disease surveillance in Indonesia through Twitter posts," *J. Appl. Res. Technol.*, vol. 18, no. 3, Jun. 2020, doi: 10.22201/icat.24486736e.2020.18.3.1091.

[11] Q. B. Baker, F. Shatnawi, S. Rawashdeh, M. Al-Smadi, and Y. Jararweh, "Detecting epidemic diseases using sentiment analysis of arabic Twits," *J. Univers. Comput. Sci.*, vol. 26, no. 1, pp. 50–70, 2020.

[12] M. Singh, A. K. Jakhar, and S. Pandey, "Sentiment analysis on the impact of coronavirus in social life using the BERT model," *Soc. Netw. Anal. Min.*, vol. 11, no. 1, pp. 1–11, 2021, doi: 10.1007/s13278-021-00737-z.

[13] S. Lim, C. S. Tucker, and S. Kumara, “An unsupervised machine learning model for discovering latent infectious diseases using social media data,” *J. Biomed. Inform.*, vol. 66, pp. 82–94, 2017, doi: 10.1016/j.jbi.2016.12.007.

[14] S. Joshi and D. Deshpande, “Twitter Sentiment Analysis System,” *Int. J. Comput. Appl.*, vol. 180, no. 47, pp. 35–39, 2018, doi: 10.5120/ijca2018917319.

[15] J. Eka Sembodo, E. Budi Setiawan, and Z. Abdurahman Baizal, “Data Crawling Otomatis pada Twitter,” 2016, doi: 10.21108/indosc.2016.111.

[16] A. Salsabila and T. Suhardijanto, “Sentiment Analysis on Indonesian Political Hoaxes,” vol. 453, no. Inusharts 2019, pp. 15–21, 2020, doi: 10.2991/assehr.k.200729.004.

[17] R. K. Putri and M. Athoillah, “Support Vector Machine Untuk Identifikasi Berita Hoax Terkait Virus Corona (Covid19),” *J. Inform. J. ...*, vol. 6, no. 3, pp. 162–167, 2021.

[18] W. H. Silitonga and J. I. Sihotang, “Analisis Sentimen Pemilihan Presiden Indonesia Tahun 2019 Di Twitter Berdasarkan Geolocation Menggunakan Metode Naïve Bayesian Classification.” Jurnal TeIKa, Vol. 9, No. 2, Oktober, 2019.

[19] T. W. Wibowo, A. F. Bustomi, and A. V. Sukamdi, “Tourist Attraction Popularity Mapping based on Geotagged Twits,” *Forum Geogr.*, vol. 33, no. 1, 2019, doi: 10.23917/forgeo.v33i1.8021.

[20] M. Gritta, M. T. Pilehvar, N. Limsopatham, and N. Collier, “What’s missing in geographical parsing?,” *Lang. Resour. Eval.*, vol. 52, no. 2, 2018, doi: 10.1007/s10579-017-9385-8.

[21] A. Dewandaru, D. H. Widyantoro, and S. Akbar, “Event geoparser with pseudo-location entity identification and numerical argument extraction implementation and evaluation in indonesian news domain,” *ISPRS Int. J. Geo-Information*, vol. 9, no. 12, 2020, doi: 10.3390/ijgi9120712.

[22] I. Zulfa, “Sistem Pemantau Influenza Like Illness Dan Visualisasinya Memanfaatkan Twitter,” Universitas Pendidikan Indonesia, 2015.

[23] R. Ranovan, A. Doewes, and R. Saptono, “Twitter data classification using multinomial naive bayes for tropical diseases mapping in Indonesia,” *J. Telecommun. Electron. Comput. Eng.*, vol. 10, no. 2–4, pp. 155–159, 2018.

[24] A. Dewandaru, S. I. Supriana, and S. Akbar, “Evaluation on geospatial information extraction and retrieval: Mining thematic maps from web source,” 2015, doi: 10.1109/ICoICT.2015.7231437.

[25] I. Zulfa and E. Winarko, “Sentimen Analisis Twit Berbahasa Indonesia Dengan Deep Belief Network,” *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 11, no. 2, p. 187, 2017, doi: 10.22146/ijccs.24716.

[26] B. Yang and T. Mitchell, “Joint extraction of events and entities within a document context,” 2016, doi: 10.18653/v1/n16-1033.

[27] J. Gelernter and S. Balaji, “An algorithm for local geoparsing of microtext,” *Geoinformatica*, vol. 17, no. 4, 2013, doi: 10.1007/s10707-012-0173-8.

[28] M. L. Khodra, “Event extraction on Indonesian news article using multiclass categorization,” 2015, doi: 10.1109/ICAICTA.2015.7335365.

[29] A. W. G. Zulkifli, “Pembobotan Fitur Ekstraksi Pada Peringkasan Teks Bahasa Indonesia Menggunakan Algoritma Genetika.” ISSN : 2355-9365, e-Proceeding of Engineering : Vol.2, No.2, Agustus, 2015, Page 6481.

[30] A. Dewandaru, S. I. Supriana, and S. Akbar, “Event-Oriented Map Extraction From Web News Portal : Binary Map Case Study on Diphteria Outbreak and Flood in Jakarta,” 2018, doi: 10.1109/ICAICTA.2018.8541345.

[31] A. El Haddadi, A. Fennan, A. El Haddadi, Z. Boulouard, and L. Koutti, “Mining unstructured data for a competitive intelligence system XEW,” in *SIIE 2015 - 6th International Conference on “Information Systems and Economic Intelligence,”* Feb. 2015, pp. 146–149, doi: 10.1109/ISEI.2015.7358737.

[32] N. Cao and W. Cui, *Introduction to Text Visualization*. doi:10.2991/978-94-6239-186-4, 2016.

[33] Di. Baviskar, S. Ahirrao, and K. Kotecha, “Multi-Layout Unstructured Invoice Documents Data set: A Data set for Template-Free Invoice Processing and Its Evaluation Using AI Approaches,” *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3096739.

[34] F. Halper, “Text Analytics Hits the Mainstream.,” *Bus. Intell. J.*, vol. 18, no. 2, 2013.

[35] R. Talib, M. Kashif, S. Ayesha, and F. Fatima, “Text Mining: Techniques, Applications and Issues,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 11, 2016, doi: 10.14569/ijacsa.2016.071153.

[36] G. Miner, D. Delen, J. Elder, A. Fast, T. Hill and R. Nisbet, “The Seven Practice Areas of Text Analytics,” in *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, doi: 10.1016/B978-0-12-386979-1.00002-5, 2012.

[37] L. Kumar and P. K. Bhatia, *Text Mining: Concepts, Process, and Applications*, vol. 4, no. 3. 2013.

[38] R. Feldman and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*, vol. 44, no. 10. 2007.

[39] K. F. Wong, W. Li, R. Xu, and Z. S. Zhang, “Introduction to Chinese natural language processing,” *Synth. Lect. Hum. Lang. Technol.*, vol. 2, no. 1, 2010, doi: 10.2200/S00211ED1V01Y200909HLT004.

[40] N. P. Katariya and M. S. Chaudhari, “Text Preprocessing for Text Mining Using Side Information,” *Int. J. Comput. Sci. Mob. Appl.*, vol. 3, 2015.

[41] V. Singh and B. Saini, “An Effective Tokenization Algorithm for Information Retrieval Systems,” 2014, doi: 10.5121/csit.2014.4910.

[42] C. Ramisch, *Multiword Expressions Acquisition: A Generic and Open Framework*. ISBN: 978-3-319-09207-2, 2015.

[43] M. Piotrowski, "Natural language processing for historical texts," *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 2, 2012, doi: 10.2200/S00436ED1V01Y201207HLT017.

[44] E. Dharmawan, H. Sujaini, and H. Muhardi, "Perbandingan Nilai Akurasi Terhadap Penggunaan Part of Speech Set pada Mesin Penerjemah Statistik," *J. Sist. dan Teknol. Inf.*, vol. 8, no. 3, p. 250, Jul. 2020, doi: 10.26418/justin.v8i3.39810.

[45] F. Haykal, A. A. Suryani, and S. Widowati, "Identifikasi Kata Majemuk Bahasa Indonesia." ISSN : 2355-9365 e-Proceeding of Engineering : Vol.7, No.2 Agustus 2020, Page 7935

[46] R. Patel and S. Patel, "Deep Learning for Natural Language Processing," in *Lecture Notes in Networks and Systems*, 2021, vol. 190, doi: 10.1007/978-981-16-0882-7_45.

[47] S. Landolt, T. Wambsganß, and M. Söllner, "A taxonomy for deep learning in natural language processing," in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2021, vol. 2020-January, doi: 10.24251/hicss.2021.129.

[48] U. Kholifah and Sabardila, "Analisis Kesalahan Gaya Berbahasa Pada Sosial Media Instagram dalam Caption dan Komentar," *Agustus*, vol. 15, no. 3, 2020.

[49] R. Al-Hashemi, "Text Summarization Extraction System (TSES) Using Extracted Keywords," *Int. Arab J. e-Technology*, vol. 1, no. 4, 2010.

[50] K. Jezek and J. Steinberger, "Automatic Text Summarization (The state of the art 2007 and new challenges)," *Proc. Znalosti*, 2008.

[51] Y. Kumar Meena and D. Gopalani, "Evolutionary algorithms for extractive automatic text summarization," in *Procedia Computer Science*, 2015, vol. 48, no. C, doi: 10.1016/j.procs.2015.04.177.

[52] J. G. Conrad and L. K. Branting, "Introduction to the special issue on legal text analytics," *Artificial Intelligence and Law*, vol. 26, no. 2. 2018, doi: 10.1007/s10506-018-9227-z.

[53] M. A. Zamzam, "Sistem Automatic Text Summarization Menggunakan Algoritma Textrank," *MATICS*, vol. 12, no. 2, 2020, doi: 10.18860/mat.v12i2.8372.

[54] S. Ceri, A. Bozzon, M. Brambilla, E. Della Valle, P. Fraternali, and S. Quarteroni, "An Introduction to Information Retrieval," in *Web Information Retrieval*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 3–11.

[55] Z. Chen *et al.*, "Information retrieval: a view from the Chinese IR community," *Front. Comput. Sci.*, vol. 15, no. 1, p. 151601, Feb. 2021, doi: 10.1007/s11704-020-9159-0.

[56] A. Esteva *et al.*, "COVID19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization," *npj Digit. Med.*, vol. 4, no. 1, 2021, doi: 10.1038/s41746-021-00437-0.

[57] M. Sarkar and S. Biswas, "Exploring Archives Space an Open Source Solution for

Digital Archiving," *DESIDOC J. Libr. Inf. Technol.*, vol. 40, no. 05, 2020, doi: 10.14429/djlit.40.05.16330.

[58] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[59] R. R. Larson, "Introduction to Information Retrieval," *J. Am. Soc. Inf. Sci. Technol.*, p. n/a-n/a, 2009, doi: 10.1002/asi.21234.

[60] S. Han, X. Hao, and H. Huang, "An event-extraction approach for business analysis from online Chinese news," *Electron. Commer. Res. Appl.*, vol. 28, pp. 244–260, Mar. 2018, doi: 10.1016/j.elerap.2018.02.006.

[61] R. Nagar *et al.*, "A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives," *J. Med. Internet Res.*, vol. 16, no. 10, p. e236, 2014, doi: 10.2196/jmir.3416.

[62] W. Wang, "Automated spatiotemporal and semantic information extraction for hazards," *ProQuest Diss. Theses*, no. August, 2014.

[63] A. Halterman, "Geolocating Political Events in Text," Proceedings ofthe Third Workshop on Natural Language Processing and Computational Social Science, Minneapolis, Minnesota, June 6, 2019, pages 29–39.

[64] Q. Li, H. Ji, and L. Huang, "Joint event extraction via structured prediction with global features," in *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2013, vol. 1.

[65] E. Aldana-Bobadilla, A. Molina-Villegas, I. Lopez-Arevalo, S. Reyes-Palacios, V. Muñiz-Sanchez, and J. Arreola-Trapala, "Adaptive geoparsing method for toponym recognition and resolution in unstructured text," *Remote Sens.*, vol. 12, no. 18, 2020, doi: 10.3390/RS12183041.

[66] P. Arcaini, G. Bordogna, D. Ienco, and S. Sterlacchini, "User-driven geo-temporal density-based exploration of periodic and not periodic events reported in social networks," *Inf. Sci. (Ny).*, vol. 340–341, 2016, doi: 10.1016/j.ins.2016.01.014.

[67] F. Hogenboom, F. Frasincar, U. Kaymak, and F. De Jong, "An overview of event extraction from text," in *CEUR Workshop Proceedings*, 2011, vol. 779.

[68] L. Zhan and X. Jiang, "Survey on Event Extraction Technology in Information Extraction Research Area," in *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Mar. 2019, pp. 2121–2126, doi: 10.1109/ITNEC.2019.8729158.

[69] Sunik, Boris, "The specification language TimeML," *SIGPLAN Notices, Vol. 40, Issue 5 May, https://doi.org/10.1145/1071221.1071224, 2005, pp 28–38*.

[70] L. D. Consortium, "ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, Version 6.6 2008.06.13," *Facilities*, 2008.

[71] J. D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii, "Overview of BioNLP'09

Shared Task on Event Extraction," 2009, doi: 10.3115/1572340.1572342.

[72] J. Barrachina *et al.*, "CAOVA: A car accident ontology for VANETs," 2012, doi: 10.1109/WCNC.2012.6214089.

[73] W. Wang and K. Stewart, "Spatiotemporal and semantic information extraction from Web news reports about natural hazards," *Comput. Environ. Urban Syst.*, vol. 50, 2015, doi: 10.1016/j.compenvurbsys.2014.11.001.

[74] K. Leetaru and P. A. Schrodt, "GDELT: Global Data on Events, Location and Tone, 1979-2012," *Annu. Meet. Int. Stud. Assoc.*, no. April, 2013.

[75] I. Owuor, H. H. Hochmair, and S. Cvetojevic, "Tracking Hurricane Dorian in GDELT and Twitter," *Agil. GIScience Ser.*, vol. 1, 2020, doi: 10.5194/agile-giss-1-19-2020.

[76] B. W. Silverman, *Density estimation: For statistics and data analysis*. 2018.

[77] D. Ahn, "The stages of event extraction," in *Proceedings of the Workshop on Annotating and Reasoning about Time and Events - ARTE '06*, 2006, pp. 1–8, doi: 10.3115/1629235.1629236.

[78] A. Judea and M. Strube, "Event extraction as frame-semantic parsing," 2015, doi: 10.18653/v1/s15-1018.

[79] J. Yang, S. Liang, and Y. Zhang, "Design challenges and misconceptions in neural sequence labeling," International Conference on Computational Linguistics, Proceedings, 2018.

[80] A. Fathan Hidayatullah and A. Sn, "ISSN: 1979-2328 UPN "Veteran," *Semin. Nas. Inform.*, vol. 2014, no. semnasIF, pp. 115–122, 2014.

[81] S. Hochreiter and J. Schmidhuber, "Long Short Term Memory. Neural Computation," *Neural Comput.*, vol. 9, no. 8, 1997.

[82] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 2012, vol. 1, doi: 10.21437/interspeech.2012-65.

[83] J. C. W. Lin, Y. Shao, J. Zhang, and U. Yun, "Enhanced sequence labeling based on latent variable conditional random fields," *Neurocomputing*, vol. 403, 2020, doi: 10.1016/j.neucom.2020.04.102.

[84] J. Hammerton, "Named Entity Recognition with Long Short-Term Memory," 2003, doi: 10.3115/1119176.1119202.

[85] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, 2011.

[86] R. Collobert, L. Bottou, J. Weston, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural Language Processing (Almost) from Scratch Ronan," *Proc. - 2017 IEEE 3rd Int. Conf. Collab. Internet Comput. CIC 2017*, vol. 2017-Janua, 2017.

[87] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," 2016, doi: 10.18653/v1/n16-1030.

[88] M. A. Murtaugh, B. S. Gibson, D. Redd, and Q. Zeng-Treitler, "Regular expression-based learning to extract bodyweight values from clinical notes," *J. Biomed. Inform.*, vol. 54, 2015, doi: 10.1016/j.jbi.2015.02.009.

[89] L. Araujo, "Genetic programming for natural language processing," *Genet. Program. Evolvable Mach.*, vol. 21, no. 1–2, 2020, doi: 10.1007/s10710-019-09361-5.

[90] R. Salgotra, M. Gandomi, and A. H. Gandomi, "Time Series Analysis and Forecast of the COVID19 Pandemic in India using Genetic Programming," *Chaos, Solitons and Fractals*, vol. 138, 2020, doi: 10.1016/j.chaos.2020.109945.

[91] B. Kranthikumar and R. L. Velusamy, "SQL injection detection using REGEX classifier," *J. Xi'an Univ. Archit. Technol.*, vol. Volume XII, no. Issue VI, 2020.

[92] J. D. Curuksu, "Principles of Data Science: Advanced," doi: 10.1007/978-3-319-70229-2_7 , 2018.

[93] B. Yu, "Three principles of data science: predictability, computability, and stability (PCS)," doi: 10.1109/bigdata.2018.8622080, 2019.

[94] C. Brando and F. Frontini, "Semantic Historical Gazetteers and Related NLP and Korpus Linguistics Applications," *J. Map Geogr. Libr.*, vol. 13, no. 1, pp. 1–6, Jan. 2017, doi: 10.1080/15420353.2017.1307307.

[95] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 1st International Conference on Learning Representations, Proceedings, 2013.

[96] A. Jaffe, Y. Kluger, O. Lindenbaum, J. Patsenker, E. Peterfreund, and S. Steinerberger, "The Spectral Underpinning of word2vec," *Front. Appl. Math. Stat.*, vol. 6, 2020, doi: 10.3389/fams.2020.593406.

[97] P. F. Muhammad, R. Kusumaningrum, and A. Wibowo, "Sentiment Analysis Using Word2vec and Long Short-Term Memory (LSTM) for Indonesian Hotel Reviews," in *Procedia Computer Science*, 2021, vol. 179, doi: 10.1016/j.procs.2021.01.061.

[98] B. Jang, I. Kim, and J. W. Kim, "Word2vec convolutional neural networks for classification of news articles and Twits," *PLoS One*, vol. 14, no. 8, 2019, doi: 10.1371/journal.pone.0220976.

[99] B. Li, A. Drozd, Y. Guo, T. Liu, S. Matsuoka, and X. Du, "Scaling Word2Vec on Big Korpus," *Data Sci. Eng.*, vol. 4, no. 2, 2019, doi: 10.1007/s41019-019-0096-6.

[100] A. Handler, "An empirical study of semantic similarity in WordNet and Word2Vec," *University of New Orleans*, Thesis, 2014.

[101] L. De Vine, G. Zuccon, B. Koopman, L. Sitbon, and P. Bruza, "Medical semantic similarity with a neural language model," 2014, doi: 10.1145/2661829.2661974.

[102] C. Cherry, H. Guo, and C. Dai, “NRC: Infused Phrase Vectors for Named Entity Recognition in Twitter,” 2015, doi: 10.18653/v1/w15-4307.

[103] T. Mcenery, V. Brezina, D. Gablasova, and J. Banerjee, “Korpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use,” *Annual Review of Applied Linguistics*, vol. 39. 2019, doi: 10.1017/S0267190519000096.

[104] N. Nesselhauf, “Korpus Linguistics: A Practical Introduction,” *Univ. Heidelb.*, vol. 2005, no. October 2005, 2011.

[105] A. M. S. Al-Hamzi, A. Gougui, Y. Sari Amalia, and T. Suhardijanto, “Korpus Linguistics and Korpus-Based Research and Its Implication in Applied Linguistics: A Systematic Review,” *Parol. J. Linguist. Educ.*, vol. 10, no. 2, 2020, doi: 10.14710/parole.v10i2.176-181.

[106] E. Tognini-Bonelli, “Korpus Linguistics at Work,” *Comput. Linguist.*, vol. 28, no. 4, 2002, doi: 10.1162/coli.2002.28.4.583a.

[107]Ng, Raymond W. M. Kwan, Alvin C.M. Lee, Tan Hain, Thomas, “Shefce: A Cantonese-English bilingual speech korpus for pronunciation assessment,” 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), doi: 10.1109/ICASSP.2017.7953273, 2017.

[108] O. Magidor, “The Philosophy of Generative Linguistics By Peter Ludlow,” *Analysis*, vol. 72, no. 4, 2012, doi: 10.1093/analys/ans095.

[109] D. Crystal, "*The Cambridge Encyclopedia of the English Language*" *Cambridge University Press*, doi*:* 10.1017/9781108528931, 2018.

[110] H. Chuquet, “Review of Altenberg & Granger (2002): Lexis in Contrast. Korpus-based Approaches,” *Lang. Contrast*, vol. 4, no. 2, 2004, doi: 10.1075/lic.4.2.09chu.

[111] M. L. Murphy, “Lexis in Contrast: Korpus-Based Approaches (review),” *Language (Baltim).*, vol. 81, no. 3, 2005, doi: 10.1353/lan.2005.0143.

[112] C. Câmpeanu and N. Santean, “On the intersection of regex languages with regular languages,” *Theor. Comput. Sci.*, vol. 410, no. 24–25, pp. 2336–2344, May 2009, doi: 10.1016/j.tcs.2009.02.022.

[113] D. Redd, B. Gibson, M. A. Murtaugh, J. Goulet, and Q. Zeng-Treitler, “Extract clinical measurement values using a regular expression pattern discovery algorithm vs support vector machine,” 2018.

[114] Y. Kim, “Convolutional neural networks for sentence classification,” 2014, doi: 10.3115/v1/d14-1181.

[115] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, “1D convolutional neural networks and applications: A survey,” *Mech. Syst. Signal Process.*, vol. 151, p. 107398, 2021, doi: 10.1016/j.ymssp.2020.107398.

[116] S. Chen, J. Yu, and S. Wang, “One-dimensional convolutional auto-encoder-based feature learning for fault diagnosis of multivariate processes,” *J. Process Control*, vol.

87, pp. 54–67, 2020, doi: 10.1016/j.jprocont.2020.01.004.

[117] G. Pavoni *et al.*, "TagLab: AI-assisted annotation for the fast and accurate semantic segmentation of coral reef orthoimages," *J. F. Robot.*, vol. 39, no. 3, 2022, doi: 10.1002/rob.22049.

[118] Fathoni, Erwin, and Abdiansah, "Multilabel sentiment analysis for classification of the spread of COVID19 in Indonesia using machine learning," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 31, no. 2, pp. 968–978, 2023, doi: 10.11591/ijeecs.v31.i2.pp968-978.

[119] I. D. Candra Arifah, "Job Replacement Artifical Intelligence Di Industri Jasa:Tinjauan Pustaka Sistematis," *J. Ilmu Manaj.* , vol. 10, no. 3, pp. 911–929, 2022.

[120] L. Li, M. Huang, Y. Liu, S. Qian, and X. He, "Contextual label sensitive gated network for biomedical event trigger extraction," *J. Biomed. Inform.*, vol. 95, 2019, doi: 10.1016/j.jbi.2019.103221.

[121] Y. Chen, "A transfer learning model with multi-source domains for biomedical event trigger extraction," *BMC Genomics*, vol. 22, no. 1, 2021, doi: 10.1186/s12864-020-07315-1.

[122] Y. Diao *et al.*, "FBSN: A hybrid fine-grained neural network for biomedical event trigger identification," *Neurocomputing*, vol. 381, 2020, doi: 10.1016/j.neucom.2019.09.042.

[123] J. Xu and M. Sun, "DPNPED: Dynamic Perception Network for Polysemous Event Trigger Detection," *IEEE Access*, vol. 10, pp. 104801–104810, 2022, doi: 10.1109/ACCESS.2022.3210697.

[124] S. K. Sahoo, S. Saha, A. Ekbal, and P. Bhattacharyya, "Event-Argument Linking in Disaster Domain," *IEEE Access*, 2022, doi: 10.1109/ACCESS.2022.3197648.

[125] Y. Zhang *et al.*, "A Question Answering-Based Framework for One-Step Event Argument Extraction," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.2985126.

[126] R. Hu, H. Liu, and H. Zhou, "Role Knowledge Prompting for Document-Level Event Argument Extraction," *Appl. Sci.*, vol. 13, no. 5, 2023, doi: 10.3390/app13053041.

[127] J. Liu, Y. Chen, and J. Xu, "Document-level event argument linking as machine reading comprehension," *Neurocomputing*, vol. 488, 2022, doi: 10.1016/j.neucom.2022.03.016.

[128] H. Wang, T. Zhu, M. Wang, G. Zhang, and W. Chen, "A prior information enhanced extraction framework for document-level financial event extraction," *Data Intell.*, vol. 3, no. 3, 2021, doi: 10.1162/dint_a_00103.

[129] Fathoni, Erwin, and Abdiansah, "Extraction of Event Sentence Information in the Covid19 Distribution Location Detection System based on the Indonesian Language Korpus," *Int. Conf. Electr. Eng. Comput. Sci. Informatics*, vol. 2022-Octob, no. October, pp. 383–388, 2022, doi: 10.23919/EECSI56542.2022.9946530.