Extraction Of Event Sentence Information In The Covid-19 Distribution Location Detection System Based On The Indonesian Language Corpus

by Fathoni Fathoni

Submission date: 07-May-2023 11:19AM (UTC+0700)
Submission ID: 2086236386
File name: ion_Detection_System_based_on_the_Indonesian_Language_Corpus.pdf (241.1K)
Word count: 5982
Character count: 32163

Extraction of Event Sentence Information in the Covid-19 Distribution Location Detection System based on the Indonesian Language Corpus

Fathoni
Department of Information System
Faculty of Computer Science,
Universitas Sriwijaya
Indralaya, Indonesia
fathoni@unsri.ac.id

Erwin
Department of Computer Engineering
Faculty of Computer Science,
Universitas Sriwijaya
Indralaya, Indonesia
erwin@unsri.ac.id

Abdiansah
Department of Informatic Engineering
Faculty of Computer Science,
Universitas Sriwijaya
Indralaya, Indonesia
abdiansah@unsri.ac.id

***Abstract*— The procedure for obtaining data on people affected by the COVID-19 disease is carried out through a manual data collection mechanism through health centers, government and private hospitals, health clinics spread throughout Indonesia, and rapid tests carried out at certain times and locations. Such a surveillance system requires time, a lot of health personnel, and is expensive. In addition, the geographical condition of the Indonesian state, which consists of many large and small islands and the vast territory of Indonesia requires another strategy to find out and make it easier to complete data on people affected by COVID-19, such as the use of information technology. The use of the Twitter dataset to detect the spread of disease in a region or country has been widely carried out by researchers. Sentence extraction is a process that must be done to facilitate the analysis of a short or very long news sentence to get the meaning and essence of the news contained in the sentence. The primary information can be identified based on keywords generated using extraction and abstraction techniques. The initial stage of the research focused on building a corpus of twitter data and a corpus of vocabulary. The first process that will be carried out is to collect natural language datasets from Indonesian-language Tweets on Twitter. Next, carry out the process of extracting incident sentence information with steps, namely making standard sentence formats, simplifying sentences, identifying essensial words, and determining input and target words in sentences.**

Keywords*— *Corpus, Covid-19, Event Extraction, Indonesian Language

I. INTRODUCTION

The increasing need for information on the location of an event causes research opportunities in the field of processing unstructured text data from data sources other than online news, such as from social media webs (such as Twitter, Facebook and Instagram). In contrast to online news information that has gone through the editorial stage to ensure the truth of the news that will be displayed on a web, the information conveyed on social media still needs to be questioned about the truth of the news. This is interesting and is an offer for the latest research to be carried out, so that information on the location of an event that is conveyed on social media can be trusted.

The latest research to determine the location of events that are sourced from unstructured data containing the name of the location must be able to process the information contained in the news document and produce geographic information in the form of coordinates or polygons that are better and more accurate [1],[2],[3],[4], but the dataset processed by the researchers was sourced from news information that has been confirmed to be true (news on line).

The use of information from Twitter as a dataset to detect the spread of disease in a region or country has been widely carried out by researchers in the world [5],[6],[7],[8]. To obtain the required dataset on Twitter, the researchers conducted data mining by utilizing the functions provided by Twitter and developing their own new additional functions that were tailored to the characteristics and topics of their respective research. Based on the published research results, the mining dataset on Twitter developed by the researchers has a level of accuracy that can still be improved, especially for the results of mining twitter data for the spread of the COVID-19 disease in Indonesia.

A lot of research has been done to find out and get the location of an event in various information in a sentence, but the research is still focused on finding words that contain the meaning of the name of the location of the incident only [9][10]. While research that discusses determining the location of the incident based on information about the incident has not been widely carried out [11][2][12] and this provides an opportunity recent research to offer models and methods that are sufficient or better to find an incident location based on sentence information written in online media. The most basic and challenging problem for research to determine the location of the event of the spread of the Covid-19 disease is the very high level of complexity which is the interaction and mixing of various aspects such as; unstructured sentence structure of tweets, characteristics and uniqueness of information sentences in tweeters as well as temporal and spatial aspects.

The complexity of the problem will be further complicated by the fact that there is a truth factor or misinformation regarding the occurrence of the occurrence of the spread of covid that was tweeted by the public on Tweeter. This provides an opportunity to find a new model that can extract information in tweets into information that

can be validated as well as detect the location of the occurrence of disease spread events (such as Covid-19) built based on datasets or corpus in Indonesian. Quick identification of the location of the spread of the COVID-19 disease is very important to be able to assist the Indonesian Government in early detection, preventing the spread and taking various strategic actions to overcome the Covid-19 pandemic in Indonesia.

This research was conducted to develop a new model or method that can extract and validate information on incident sentences and capture data on the distribution and location of the spread of the COVID-19 disease in Indonesia with a better level of accuracy than previous researchers. Research using machine learning to build a model that can classify and identify tweets that indicate the spread of disease in Indonesia has been carried out by several previous researchers [13],[14],[5]. This study resulted in a fairly good classification of disease tweets, but did not convey information on the type and location of the spread of the disease, as well as the level of accuracy that could still be improved. The use of machine learning in this study is also intended to build a classification model and grouping the results of identifying tweet data on the spread of COVID-19 in Indonesia using a method that has a better accuracy rate than previous research. To achieve a better level of accuracy, machine learning algorithms will be used that are in accordance with the characteristics and types of datasets generated in the early stages of this research.

II. RELATED WORK

A. State of the Art

The latest research to determine the location of events originating from unstructured data containing the name of the location of the event must be able to process the meaning of the information contained in the news document and produce geographic information in the form of coordinates or polygons that are better and more accurate [2],[3],[4],[9], but the dataset processed by the researchers was sourced from news information that had gone through the editorial stage so that it could be confirmed (online news).

Many studies have been carried out to find out and get the location of an event in various information in a sentence, but the research is still focused on finding words that contain the meaning of the name of the location of the incident only [10],[15]. While research that discusses determining the location of events based on information on events has not been widely carried out [3],[4],[14] and this condition provides the latest research opportunities to offer models and methods that are sufficient or it is better to find an incident location based on sentence information written in online media. The most basic and challenging problem for research to determine the location of the event of the spread of the Covid-19 disease is the very high level of complexity which is the interaction and mixing of various aspects such as; the unstructured sentence structure of the tweet, the characteristics and uniqueness of the informational sentence on the tweeter as well as the temporal and spatial aspects as well as the certainty of the truth of the information conveyed on social media which is not necessarily in accordance with the conditions at the scene.

Research using machine learning to build a model that can classify and identify tweets that indicate the spread of disease in Indonesia has been carried out by several previous researchers [5],[16],[17]. This study resulted in a fairly good classification of disease tweets, but did not convey information on the type and location of the spread of the disease, as well as the level of accuracy that could still be improved.

Search for the location of the incident [2] which tries to resolve the main geographic focus of an event in political news documents from various online media sources by using sentence information using word embedding Word2Vec. The performance of F-1 from determining this location in carrying out resolutions reached 82.90% for the corpus and training set from the same news agency. However, when faced with a corpus of various agencies, the accuracy drops to 64.21%. The limitation of this method is that only an event and a location can be returned from a document. This makes it difficult to apply in many cases when there are several events or several locations at the same time being the focus of an event.

Another search location research, namely Geoparser and Geocoder Camcoder works [4] distinguishes between literal toponyms and associative toponyms as an effort to geolocate events. The method used is an artificial neural network with an NCRF++ geotagger and achieves an F-Score accuracy of 77.6% for the identification of fine-grained toponyms. Meanwhile, the Mordecai geoparser uses LSTM, Glove's semantic information feature, and uses event triggers to find the location of events with an F-Score of 84%.

Event extraction is a problem in information extraction which aims to decompose text into event structures such as event triggers, semantic roles and event arguments related to these semantic roles ([18]. Here the definition of the event in question is an event that involves several arguments. This definition does not require the extraction of the time argument (although the time argument is also used in the corpus and the proposed model, but the temporal aspect is not the main discussion). The integration of event extraction into event location search can be done by integrating event extraction stages such as trigger identification, identification and assignment arguments, and event correlation into regular event location search workflows consisting of geotagging and geocoding. This integration is felt to be able to answer the problem of improving the quality and performance of geoparsing by providing a semantic context of events in the inference process.

Both regular event location searches (with toponym resolution scope and document resolution scope) and event location searches in general are still dominated by geoparsers operating in the English domain. In state-of-the-art NLP methods that increasingly rely on machine learning, the search for event locations is of course trained using a dataset or corpus from a specially defined language domain, so that geoparsing applications to other languages are constrained, because the inference model with different languages generally won't work well. Even in the same language, application in different domains is also a challenge. For example, an incident location search model that focuses and is trained with a microtext domain (informal and short such as Twitter messages) such as [12] will not be optimal and suitable if it is used for case prediction in the news domain (formal and has a certain editorial process).

This limitation shows the importance of researching an incident location search model that is focused and built with

the Indonesian language corpus, which can be started in certain domains, such as the realm of information (news) delivered on social media (twitter) as is the scope of this dissertation. One of the obstacles in the development of this model is the absence of a corpus and a prototype for finding the location of the incident in Indonesian [19]. The prototype in Indonesian related to finding the location of the incident in general is just a text geolocation method from twitter which is generally done with the Twitter API using a GPS signal recorded by the device when the tweet is made, for example [20],[21]. This has a weakness where the location conveyed by the GPS signal is not necessarily the same as the location of the incident conveyed in the tweet sentence. There is no method or prototype of a geoparser with the aim of geolocation of unstructured text. Perhaps the closest work is a prototype of the 5W1H event-extraction model [19] which only stops at geotagging (in determining the where tag) and does not have the purpose of disambiguation and resolution.

B. Electronic Data Source

Changes in people's behavior in the era of information technology which brought about very significant changes to the ways and techniques of disseminating news which were originally dominated by print, radio and television media have now switched to using internet media. This increase in the spread of news on the internet is marked by the large number of web pages, namely 170 Terabyte pages, approximately 1,354,440 GB of audio-visual data on social media and at least 5,000 more databases in the form of scientific papers, economic data, patents and others. Massive increase in the spread of news occurs in social media such as Twitter with 342,000 tweets per minute, 3,298,560 posts on Facebook per minute and 276,480 searches on Google per minute [22].

Twitter is one of the social media that is growing rapidly with an increasing number of users throughout the world, including in Indonesia. Basically, tweets on Twitter are text. Text is a sentence of information and knowledge that is disseminated at a certain time and medium [23]. In the current information era, there is a lot of information available on the internet in the form of text from various types of documents such as research documents, magazines, electronic books and articles that are unstructured (unstructured data), including emails, pdf files, social media (twitter, facebook). and others), video, audio, image and business content in bulk [24]. The volume of text documents is growing rapidly and experts predict a growth of 80% by 2025 [25]. Search Technologies states that 80% of data in organizations is unstructured data.

III. METHODOLOGY

The research stages starting from the early stages of research focused on building the twitter data corpus and vocabulary corpus. The first process that will be carried out is to collect natural language datasets from Indonesian-language Tweets on Twitter. Next, carry out the process of extracting incident sentence information with steps, namely making standard sentence formats, simplifying sentences, identifying important words in sentences, and determining input and target words in sentences.

The second stage of research starts from data validation on the main corpus for twitter data and the vocabulary corpus by identifying input and target keywords and identifying the location of the incident words in the vocabulary. The next step is to identify and capture the location of the incident by performing the following steps, namely identifying the entity, Location means the location of the incident, the process of extracting sentences meaning the location of the incident and identifying the location of the pseudo-event and the coordinates of a valid geographic map. The last stage is to identify the location of the incident in the form of location attributes, number, location coordinates and accuracy values.

The stage of identification and capture of the location of the incident is carried out by making a model based on the identification of needs and the final results to be achieved. The proposed model has three main stages and is distributed into six stages, as shown in Fig. 1

The first stage is Identification of Entity Meaning Location of Event, which is divided into two processes, namely:

1. The process of arresting and identifying entities that contain the meaning of the word location of the incident (spread of covid-19 disease)
2. The process of identifying pseudo-event locations and valid geographic maps

These two stages aim to facilitate access to meaningful entity information such as the location of the incident and create better integration.

The second stage is the process of extracting sentences with the meaning of events, which is divided into two processes, namely:

1. The process of identifying and extracting Event Triggers (events/events) and classifying them into appropriate labels based on the event code (events)
2. The process of identifying and extracting Event Arguments contained in news texts containing the meaning of events (the spread of COVID-19) and classifying them into different event arguments.

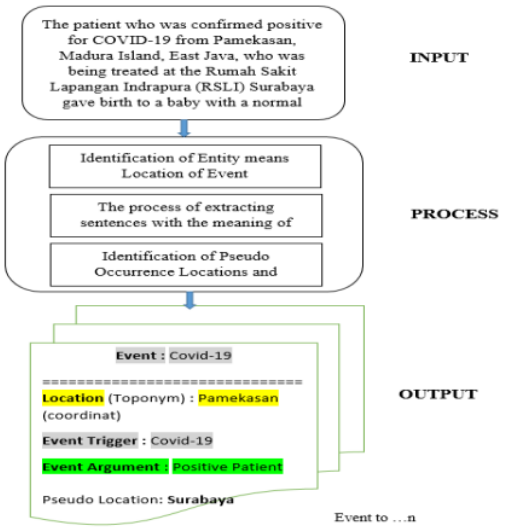

Fig. 1. Flowchart of the Stages of Identification and Capture of Event Locations.

IV. RESULTS AND DISCUSSION

A. Main Corpus and Vocabulary Corpus

Main Corpus is a Corpus that contains all text information in the form of a dataset in the form of a standard file containing important notes (annotations) of an entity and which will be processed and separated by separator code. The main corpus in this study is a research dataset that was obtained from the stylistic speech contained in the twitter user's tweets which were identified as containing the meaning of sentences containing elements of the location of the occurrence of the spread of the covid-19 virus, which is presented in Table 1.

TABLE 1. Main Corpus example of a user's tweet on Twitter showing the location

No	Contents of Information submitted
1	Tau nggak sih kenapa banyak masyarakat yang mau di vaksin? ""Biar virus corona hilang dari tangerang"" salah dong, mereka vaksin biar bisa berlibur dan dapet diskon. Coba liat mall sudah rame, banyak diskon juga kalau nunjukin sertifikat vaksin. Gila …..
2	waktu denger NF kena corona dan dikarantina dibandung, aku khawatir. Tapi, di satu sisi aku yakin kalian bisa menghadapinya, dan waktu aku denger kalian sembuh, aku senang, kalian membuktikannya!
3	Kasus aktif Corona di DKI Jakarta semakin berkurang dari hari ke hari. Rumah sakit rujukan COVID-19 juga akan mulai kembali menerima pasien non-COVID
4	Makin Membaik, Covid-19 di Sragen Hanya Tambah 10 Kasus dan Pasien di ICU Tinggal Satu Orang Hari Ini
5	Kasus Aktif Covid-19 di Kota Bandung Totalnya 364, Pasien Sembuh Tembus 40 Ribu Orang
6	155 Pasien Covid-19 Diisolasi di 5 Fasilitas Isoter di Pekanbaru
7	Empat Ribu Pasien Covid-19 di Aceh Masih Dirawat
8	Pasien terkonfirmasi positif COVID-19 asal Pamekasan, Pulau Madura, Jawa Timur, yang sedang dirawat di Rumah Sakit Lapangan Indrapura (RSLI) Surabaya melahirkan bayi dengan persalinan normal.
9	Bayi yang lahir dari pasien Covid-19 di RSLI Surabaya berjenis kelamin laki-laki dengan berat 2.500 gram
10	JUMLAH pasien positif Covid-19 di ruangan isolasi RSUD TC.Hillers Maumere terus menurun seiring dengan menurunnya kasus Covid-19 di Kabupaten Sikka, Nusa Tenggara Timur. Saat ini, rumah sakit tersebut hanya merawat dua orang pasien Covid-19
11	Alhamdulillah, Jumlah Pasien Covid-19 yang Dirawat di RS Garut Tinggal 7 Orang

The news text contained in the main corpus is planned to be broken down into several sentences and separated into several token levels with the basic format as shown in Table 2.

TABLE 2. Basic Format of Annotations on Main Corpus

No	Format Type	Format
1	Standard	<token>/<POS Tag>/<tag entitas>/<…>
2	LOCATION OF THE INCIDENT (LOKJ)	<token>/<POS Tag>/<LOKJ
3	Event Trigger (EVT)	<token>/<POS Tag>/<EVT>/<event subtypes>
4	Argument (ARG)	<token>/<POS Tag>/<ARG>/<tag role argumen>

This stage will utilize the Part-of-Speech Annotation (POS) tagger to make it easier to identify entities in the main corpus text. To simplify the process of annotating the training dataset, it is planned to use the InaNLP tool using the Hidden Markov Model (HMM) basis for Indonesian in defining part-of-speech tags for each token (words, numbers and symbols). Furthermore, it was annotated again as a new proposal to add an Indonesian language corpus repository which would be called the Vocabulary corpus.

B. Basic Indonesian Sentence Structure

The words contained in the language can be processed and categorized into labels or word classes which are then given special assessment notes (annotations) using Part-of-speech (POS) taggers. While the word class (part of speech) is a group of words that have similarities in their formal behavior. In the context of research on natural languages, the availability of data classes is very necessary, especially for making natural language applications [26], because in this process each word (token) will be given a unique label according to the part of the speech called POS marking (POS). tagging) [27],[28]. POS tagging is one of the basic functions that is very useful for many natural language applications [29],[30].

The use of word labels in the Kamus Besar Bahasa Indonesia(KBBI) consists of seven groups of word labels, namely: a (adjective), adv (adverb), n (noun), num (numeral), p (particle), pron (pronoun), and v (verb). While in English there are eight groups of word labels, namely: Noun (N), Verb (V), Adjective (A or Adj), Adverb (Adv), Preposition (P), Article or Determiner (Det), Conjunction (Conj) and Interjection (Int). There are 26 word labels in Indonesian as shown in Table 3.

This means that an incomplete language cannot be called a sentence, but only in the form of words or groups of words. Another feature of speech referred to as a sentence is that there is a predicate function in it. The core structure of Indonesian is characterized by the presence of a subject and predicate function (S-P), and can be expanded with an object (O), complement (Pel), and/or adverb (K), described as S + P + ({O} + {Pel} + {K}).

A good sentence with a complete syntax function is a very effective communication medium in human life. But the fact is that in the use of everyday sentences there are still errors in the use of Indonesian, especially in the delivery of information on social media. These errors can be in the form of: (1) spelling errors (2) errors in the use of diction (3) incorrect grammatical structures (5) combinations of the use of Indonesian and foreign languages simultaneously, and (6) variations of Indonesian with slang, English foreign and local languages [31].

TABLE 3. Indonesian word labels

No.	Code	Word class	Information
1	NNO	*Noun*	Noun
2	NNP	*Proper Noun*	Unique individual
3	PRI	*Interogative Pronoun*	To replace the matter in question.
4	PRK	*Clitized Pronoun*	variants tied to the pronominal persona you, me and him.
5	PRN	*Pronoun*	Replace people, things /something that is objected.
6	PRR	*Relative Pronoun*	Replace the main part and/or connect it with an explanatory part.
7	ADJ	*Adjective*	Is an adjective, a word that describes a noun.
8	VBE	*Existential Verb*	is a verb existential in the sentence.
9	VBI	*Intransitive Verb*	declaring deeds, process, work, action
10	VBL	*Linking Verb*	Connect the two parts namely subject and complement
11	VBP	*Passive Verb*	Affixed passive verb*ter-* or *ke-an.*
12	VBT	*TransitiveVerb*	Examples : *membaca, menyirami.*
13	ADK	*Adverb time*	Examples :*akan, bakal, belum, lagi, masih, maupernah, sedang.*

14	ADV	*Adverb*	Explain the other words.
15	NEG	*Negation*	Examples : *tidak, bukan, kagak, tak, enggak.*
16	CCN	*Coordinative Conjuntion*	Connect two/more clauses, equivalent sentences.
17	CSN	*Subordinative Conjunction*	Connect two/more clauses, unequal sentences
18	PPO	*Preposition*	elements forming prepositional phrases, such as *di, ke,dari, untuk, oleh, tentang.*
19	INT	*Interjection*	Expressing feelings
20	KUA	*Quantifier*	Examples : *sesuatu, semua, beberapa, sebagian.*
21	NUM	*Numeral*	To count things.
22	ART	*Articles*	limiting nouns,like *sang, si, kaum.*
23	PAR	*Particle*	Firms or fillers, such as *pun, per.*
24	UNS	*Unit Symbols*	Examples : *W, kg, km, meter.*
25	$$$	*Currency*	Examples: *$, Rp..*
26	SYM	*Character Symbols*	Examples : *?, !, -, α, Σ.*

C. Sentence and Event Extract

Sentence extraction is a process that must be done to facilitate the analysis of a short or very long news sentence to get the meaning and essence of the news contained in the sentence. The main information can be identified based on key words generated using extraction and abstraction techniques. Abstraction technique is a text summarization technique that is carried out by taking important information from the source document and then processing it so as to produce a summary using new sentences that are not contained in the original document [32]. While the sentence extraction technique is a comprehensive summary model that includes word sequences in rewritten sentences and the process of selecting key words in sentences originating from original documents [33].

Event extraction is the development of research from Relationship Extraction which is limited to the relationship between two entities (such as affiliation between the actor and the actor's organization or state or the relationship between two actors in the event), detection and recognition of events is able to analyze text to find out pairs of more than two entities, such as who the perpetrator was, what the perpetrator did, who ordered the perpetrator, when it was carried out and where the incident occurred (entity mentions, event triggers, event arguments). An example will be illustrated in Fig. 2.

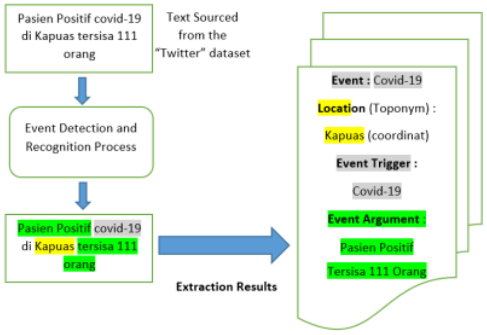

Fig. 2. Detection and introduction of events with cases of the spread of Covid-19 disease

Figure 2 informs the basic format that is formulated to extract text containing news of the spread of the Covid-19 disease at an incident location marked by the Event Trigger "Covid-19" with an Argument entity group (event) consisting of " pasien positif" and " tersisa 111 orang" and Kapuas indicating a location where the event occurred (geographical location).

The discussion of event extraction is more interesting because the events contained in the text will be analyzed for the meaning of the word (semantics) so that the same event conveyed through syntactically different language styles will produce the same conclusion output of the same event.

The probability of events that identify different speeches or styles of language conveyed by humans on online social media (such as Twitter and Instagram) that contain the same meaning of information is very high, this is because social media users are given great freedom to be allowed to write their preferred language style, as shown in Table 4.

TABLE 4. Illustration of text on social media that have the same meaning

No	Contents of Information submitted
1	Tingkat Kesembuhan Pasien Covid-19 di Bengkalis Terus Membaik
2	Di Bengkalis tingkat kesembuhan pasien Covid-19 terus membaik
3	Tingkat Kesembuhan Pasien Covid-19 Terus Membaik di Bengkalis

D. Semantic Framework of Event Extraction

Based on the understanding of "events" in event extraction research, two main models of the event concept have been used as references by many researchers, namely the TimeML model and the ACE model. The TimeML model focuses on recording the occurrence time of each event and models four main data structures, namely: EVENT, TIMEX3, SIGNAL, and LINK [34].

TimeML introduces temporal functions to express the timeframe of an event, such as: last month, yesterday and so on. TimeML annotations have a dependency structure between events with the concept of LINK, so that anchoring (event A occurs at time T), sequence (event A occurs after event B), or embedding (event A within event B) can be expressed in a detailed and unambiguous way. The TimeML model emphasizes the temporal aspect of events and does not model the spatial or grouping aspects of event arguments.

ACE only focuses on events that are interesting enough to be noticed. Extraction of the ACE model is more complex than TimeML because it involves the detection of Trigger (anchor) events, identification of arguments and their semantic roles, as well as grouping and correspondence of events. The TimeML model emphasizes the temporal aspect of events and does not model the spatial or grouping aspects of event arguments. The ACE model divides entities into seven main categories, namely person (individual / PER), organization (organization / ORG), location (location / LOKJ), geo-political entity (geopolitical entity / GPE), facility (facility / FAC), vehicle (vehicle / VEH), and weapon (weapon / WEA). In the ACE model, the concept of "event" is defined as an event that changes the state, which is indicated in the text by the occurrence of a trigger.

If the extraction system is equipped with a typical semantic framework of events as in the ACE model taxonomy, for example in Fig. 3. which informs the simple ontology reference of the CAOVA adaptation [35] then the system can reason that the accident has special semantic

roles such as what are the vehicles involved (Vehicle-Argument), the origin of a vehicle (From-Argument) and so on. can improve the quality of inference.

> *Terjadi Kecelakaan di jalan Soekarno Hatta Palembang Antara truk bermuatan ternak Sapi dari Mesuji Lampung dengan Dua sepeda motor.*

Fig. 3. Semantic illustration in ontology with CAOVA

An inference system that performs event extraction with good performance is able to assign assignments to the From-Argument slot, so it can distinguish that "Mesuji" and "Lampung" are related from From-Argument, not PlaceArgument. The implication is that the main event (Accident) inference performance is potentially better.

V. CONCLUSION

Determining the location of the event of the spread of the Covid-19 disease acquires a very high level of complexity which is the interaction and mixing of various aspects such as; unstructured sentence structure of tweets, characteristics and uniqueness of information sentences in tweeters as well as temporal and spatial aspects.

The Indonesian language corpora that was built and used to train the datasets in this study will consist of; 1) The main corpus sourced from the tweet dataset of tweeter users in Indonesia and 2) The corpus vocabulary of sentence information that has been validated and contains the meaning of the word location and events of the spread of the COVID-19 disease.

REFERENCES

[1] M. B. Imani, S. Chandra, S. Ma, L. Khan, and B. Thuraisingham, "Focus location extraction from political news reports with bias correction," in *2017 IEEE International Conference on Big Data (Big Data)*, 2017, pp. 1956–1964.

[2] 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, pp. 118–136, 2019.

[3] 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.

[4] A. Halterman, "Geolocating Political Events in Text," in *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, 2019, pp. 29–39.

[5] M. Adriani, F. Azzahro, and A. N. Hidayanto, "Disease surveillance in Indonesia through Twitter posts," *J. Appl. Res. Technol.*, vol. 18, no. 3, pp. 144–153, 2020.

[6] Q. B. Baker, F. Shatnawi, S. Rawashdeh, M. Al-Smadi, and Y. Jararweh, "Detecting Epidemic Diseases Using Sentiment Analysis of Arabic Tweets," *J. Univers. Comput. Sci.*, vol. 26, pp. 50–70, 2020.

[7] 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, p. 33, 2021.

[8] 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.

[9] W. H. Silitonga and J. I. Sihotang, "Analisis Sentimen Pemilihan Presiden Indonesia Tahun 2019 Di Twitter Berdasarkan Geolocation Menggunakan Metode Naïve Bayesian Classification," *TelKa*, vol. 9, no. 02, pp. 115–127, 2019.

[10] T. W. Wibowo, A. F. Bustomi, and A. V. Sukamdi, "Tourist Attraction Popularity Mapping based on Geotagged Tweets," *Forum Geogr.*, vol. 33, no. 1, pp. 82–100, 2019.

[11] M. Gritta, M. T. Pilehvar, N. Limsopatham, and N. Collier, "What's missing in geographical parsing?," *Lang. Resour. Eval.*, vol. 52, no. 2, 2018.

[12] 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," *IJIRS Int. J. Geo-Information*, vol. 9, no. 12, 2020.

[13] I. Zulfa, "SISTEM PEMANTAU INFLUENZA LIKE ILLNESS DAN VISUALISASINYA MEMANFAATKAN TWITTER," Universitas Pendidikan Indonesia, 2015.

[14] 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. 2–2, pp. 155–159, 2018.

[15] A. Dewandaru, S. I. Supriana, and S. Akbar, "Evaluation on geospatial information extraction and retrieval: Mining thematic maps from web source," in *2015 3rd International Conference on Information and Communication Technology, ICoICT 2015*, 2015.

[16] I. Zulfa and E. Winarko, "Sentimen Analisis Tweet Berbahasa Indonesia Dengan Deep Belief Network," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 11, no. 2, p. 187, 2017.

[17] B. Yang and T. M. Mitchell, "Joint Extraction of Events and Entities within a Document Context," in *Proceedings of the 2016 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 289–299.

[18] J. Gelernter and S. Balaji, "An algorithm for local geoparsing of microtext," *Geoinformatica*, vol. 17, no. 4, pp. 635–667, 2013.

[19] M. L. Khodra, "Event extraction on Indonesian news article using multiclass categorization," in *2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 2015, pp. 1–5.

[20] A. W. G. Zulkifli, "Pembobotan Fitur Ekstraksi Pada Peringkasan Teks Bahasa Indonesia Menggunakan Algoritma Genetika."

[21] 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," in *ICAICTA 2018 - 5th International Conference on Advanced Informatics: Concepts Theory and Applications*, 2018.

[22] 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,"* 2015, pp. 146–149.

[23] N. Cao and W. Cui, *Introduction to Text Visualization*. 2016.

[24] Di. Baviskar, S. Ahirrao, and K. Kotecha, "Multi-Layout Unstructured Invoice Documents Dataset: A Dataset for Template-Free Invoice Processing and Its Evaluation Using AI Approaches," *IEEE Access*, vol. 9, 2021.

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

[26] M. Piotrowski, "Natural language processing for historical texts," *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 2, 2012.

[27] 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.

[28] F. Haykal, A. A. Suryani, and S. Widowati, "Identifikasi Kata Majemuk Bahasa Indonesia."

[29] R. Patel and S. Patel, "Deep Learning for Natural Language Processing," in *Lecture Notes in Networks and Systems*, 2021, vol. 190.

[30] 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-Janua.

[31] U. Kholifah, D. A. Sabardila, "Analisis Kesalahan Gaya Berbahasa Pada Sosial Media Instagram dalam Caption dan Komentar," 2020.

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

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

[34] J. Pustejovsky and B. Ingria, "The specification language TimeML," *Lang. ...*, 2005.

[35] J. Barrachina *et al.*, "CAOVA: A car accident ontology for VANETs," in *IEEE Wireless Communications and Networking Conference, WCNC*, 2012.

Extraction Of Event Sentence Information In The Covid-19 Distribution Location Detection System Based On The Indonesian Language Corpus

ORIGINALITY REPORT

8%	8%	8%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Sriwijaya University Student Paper	2%
2	www.preprints.org Internet Source	1%
3	www.scielo.org.mx Internet Source	1%
4	Heyan Huang, Xiao Liu, Ge Shi, Qian Liu. "Event Extraction With Dynamic Prefix Tuning and Relevance Retrieval", IEEE Transactions on Knowledge and Data Engineering, 2023 Publication	1%
5	www.jist.ir Internet Source	1%
6	mediaindonesia.com Internet Source	1%
7	sci-hub.st Internet Source	1%

8	thesai.org Internet Source	1 %
9	Javier Osorio, Alejandro Beltran. "Enhancing the Detection of Criminal Organizations in Mexico using ML and NLP", 2020 International Joint Conference on Neural Networks (IJCNN), 2020 Publication	1 %
10	www.periodicos.ulbra.br Internet Source	1 %
11	Submitted to SVKM International School Student Paper	1 %

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	Off		