

Analysis of named-entity effect on text classification of traffic accident data using machine learning

Anugrah Dwiatmaja Putra, Abba Suganda Girsang

Department Computer Science, BINUS Graduate Program–Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Aug 8, 2021

Revised Dec 9, 2021

Accepted Jan 11, 2022

Keywords:

Classification

Machine learning

Named-entity

Social media

Traffic accident analysis

ABSTRACT

With the rising number of accidents in Indonesia, it is still necessary to evaluate and analyze accident data. The categorization of traffic accident data has been developed using word embedding, however additional work is needed to achieve better results. Several informative named entities are frequently sufficient to differentiate whether or not information on a traffic accident exists. Named-entities are informational characteristics that can offer details about a text. The influence of named-entities on thematic text categorization is examined in this paper. The information was collected using a Twitter social media crawl. Preprocessing is done at the beginning of the process to modify and delete useful text as well as label specified entities. On support vector machine (SVM), scheme comparisons were performed for i) word embedding, ii) the number of occurrences of named entities, and iii) the combination of the two is known as a hybrid. The hybrid scheme produced an improvement in classification accuracy of 90.27% when compared to word embedding scheme and occurrences of named entities scheme, according to tests conducted using 1.885 data consisting of 788 accident data and 1.067 non-accident data.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Anugrah Dwiatmaja Putra

Computer Science Department, BINUS Graduate Program–Master of Computer Science

Bina Nusantara University

Jakarta, Indonesia

Email: anugrah.putra@binus.ac.id, anugrahdputra@gmail.com

1. INTRODUCTION

In Indonesia, the frequency of accidents is extremely significant, with all types of injuries, including death. According to World Health Organization (WHO) data [1] in 2016, 31.282 people died in a total of 106.644 road accidents in Indonesia, with 78% of males and 22% of women. This implies that 12,2 persons died in a traffic accident for every 100,000 inhabitants, resulting in a mortality rate of 29.3%. In recent years, there has been a surge in the research of traffic accidents as a result of crowdsourcing data to supplement conventional techniques and uncover new facts. Twitter, which has gotten a lot of attention in recent years, is slowly becoming acknowledged as a source of information for users' direct contributions to event detection. Twitter has at least 30 million users in 2010 [2]. Twitter creates an online ecosystem in which information is generated, consumed, promoted, disseminated, discovered, and shared for particular reasons, most of which are linked to community and social activities rather than functional task-oriented goals. As a result, social media sites like Twitter will serve as data sources, and it will be possible to obtain a wide range of information from a diverse group of individuals in a timely way.

Information may be easily collected and then analyzed and categorised according to certain categories using this enormous amount of data, particularly information relating to traffic accidents such as [3]. This study

uses crawling to collect data on traffic accidents, which is then categorized into two categories: true or false on traffic accident news. Facebook's fasttext technique is used to weight word representations. The words in the document are used as quantitative characteristics in several techniques to text categorization that are based on the machine learning (ML) algorithm. The assumption behind this technique is that the frequency of particular terms in a text is a good predictor of a broad topic. This implies that named entities could be a better fit for text document categorization. The influence of named-entities on the categorization of traffic accident information data text will be investigated in this study. The comparison will be done in three ways: utilizing fundamental techniques that word embedding (Word Embedding), the numbers of occurrence named entities (Named Entities), and a mix of the two (Hybrid). The dataset from the previous study [3] will be utilized and combined with the most recent crawling dataset, which will then be labeled with named-entities. In ML, the basic algorithms to be utilized are support vector machine (SVM). The following are some of the study's contributions: i) The dataset is made up of preprocessed text from prior research datasets and new crawling methods. Furthermore, the data entity labeling is done with the help of a preset label; ii) Text categorization for traffic accident data use the SVM method, which compares predefined named entities to three predetermined schemas: word embedding, named entities, and hybrid. The data utilized in this study is the result of crawling from the social networking site Twitter, which yielded 1,885 results. The study then concentrates solely on the use of Indonesian and the previously specified set of named things.

2. RESEARCH METHOD

Web crawlers have been around almost as long as the world wide web. In 1993, the first crawler was implemented. For mining huge datasets, web crawling is used to index information on a website utilizing a uniform resource locator (URL) and an application programming interface (API). Crawlers lead to a process of document sharing, more about interactive content, and even full-fledged apps as the web advances [4]. After Facebook and Instagram, Twitter is the world's third most popular online social network (OSN), with a simple data model and direct data access API. It's therefore excellent for social network research involving hundreds of millions of people [5]. When it came to data access, Twitter used to have a fairly liberal approach [6]. Twitter began imposing tougher limitations in 2021, as stated by the official Twitter blog [7], because it was concerned that third-party services would exploit the API and develop apps that basically mimicked its primary feature. Twitter has a straightforward data delivery strategy that is supported by a highly efficient and scalable infrastructure [8]. There are numerous ways to access information from Twitter, one of which is to utilize the Twitter developer page's application program interface (API).

In numerous application domains, SVM is one of the most resilient and robust classification and regression methods. The basic goal of SVM is to use a surface that optimizes the margin between classes in the training set to separate them [9], [10]. A set of n instances is required to train an SVM. Each example is made up of two parts: an input vector x_i and a label y_i . Assume that the training set X is $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. For We'll use the example of a two-dimensional input, i.e., $x \in \mathbb{R}^2$, for illustration purposes. There are various hyperplanes that can be split, and the data can be divided linearly. The generalizability, on the other hand, is dependent on the position of the separator hyperplane and the hyperplane with the greatest margin.

A named entity is a term that denotes that an element has properties with a group of other items [11]. Entity extraction from a set of words is a method of detecting and classifying entities, also known as named entity recognition (NER). NER is significant in different natural language processing (NLP) tasks such as text interpretation, information retrieval, automatic text summarization, machine translation, and knowledge base development, in addition to the key subtask of information extraction [12]. The NER-based clustering method pulls named items from groups based on contextual similarity. The use of unlabeled data, according to Collins [13], lowers the monitoring needs to only seven basic principles.

NER is used in supervised learning to solve multi-class classification and sequence labeling problems [14]. The features in annotated data samples are meticulously constructed to reflect each training occurrence. Machine learning techniques are then used to examine the model in order to detect similar patterns in previously unseen data. In a supervised NER system, feature engineering is critical. A feature vector representation is a text abstraction in which one or more boolean, numeric, or nominal values represent a word [15]. The supervised NER has made extensive use of the word level function, list search feature, and corpus feature. Many machine learning methods have been built in the supervised NER based on these characteristics [16], [17].

Accident-related research has increased in recent years as a result of crowdsourcing data to supplement established approaches and uncover new facts. Twitter, which has gotten a lot of press in recent years, has steadily gained acceptance as a source of information for users direct contributions to event detection. There were at least 30 million Twitter users in 2010, while there were 330 million in 2019 [18]. Twitter creates an online ecosystem where information is generated, consumed, promoted, disseminated,

found, and shared for particular community and social reasons rather than task-oriented functional ones [19]. As a result, social media sites like Twitter will serve as data sources, allowing for the rapid retrieval of a wide range of information from a large number of individuals. Separating data that contains or does not contain traffic accident information requires data processing. This is because utilizing the keyword "accident" in the crawling technique will also return data that does not contain traffic accident information but has the same word component. In the paper, Saputro and Girsang [3], achieved the best accuracy of 88% in his study by categorizing utilizing the SVM approach based on FastText representation to tackle the problem.

Several informative named entities are frequently sufficient to differentiate whether or not information on a traffic accident exists. For example, information on traffic accidents will include additional details such as location, casualty injury, and time. In the meanwhile, data that does not include accident information is less likely to have several sets of such data. As a result, we believe that named entities are a feature that may be utilized to separate data into defined categories. This is due to the fact that named entities are distributed across the item response theory (IRT) hierarchy in various categories. On articles data [20], the usage of named entities in text classification was used to categorize the categories of presidential election news depending on their nation of origin, resulting in an increase in the micro average F1 score for the closest category to 81.4%.

3. PROPOSED METHOD

A hierarchical text classification aims to classify each incoming document into zero, one, or several categories in the text hierarchy. One approach to this technology, SVM with combination scheme, has demonstrated significant benefits in a variety of text categorization tasks. SVM's performance is dependent on the kernel functions and slack variables used. To put it another way, optimizing the two parameters is crucial for optimizing the SVM algorithm [21].

The steps of this research method are depicted in Figure 1. This research uses a dataset gathered from Twitter Indonesian language and keywords that correspond to "traffic accidents". To see how the named-entity impacts the social media text categorization of traffic accident information, the classification technique will be coupled with the named-entity approach as a text representation. To clear data from noise, preparation is required early on. The final stage is to assess the model to see how named-entities affect the text classification model and which model produces the best results.

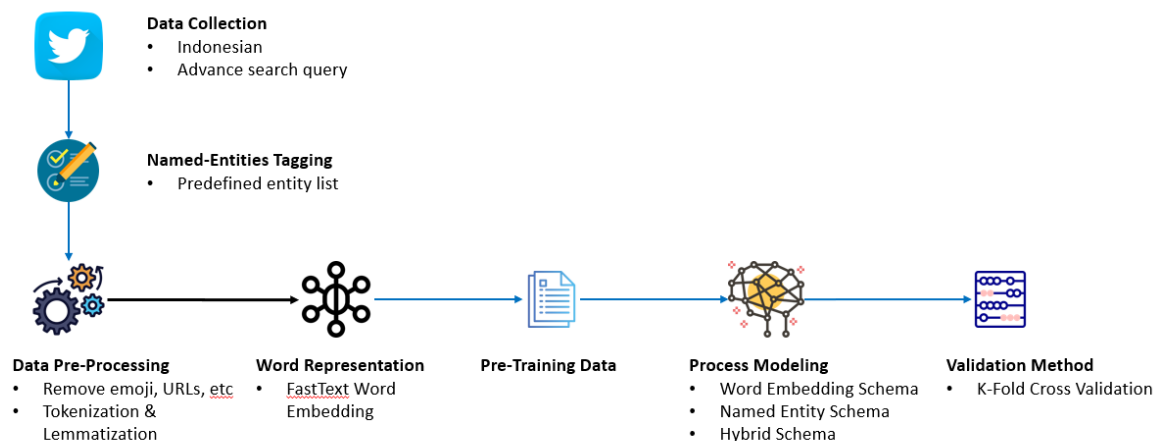


Figure 1. Proposed method

The hybrid schema is presented as a new schema that combines the word embedding and named entity schemas. As illustrated in Figure 2, the hybrid schema is constructed by integrating sentence probability evaluations against labels. The computed ratio is then used to calculate the contribution of each schema to the hybrid schema, ensuring that the contributions are balanced and that the data prediction findings are as accurate as possible.

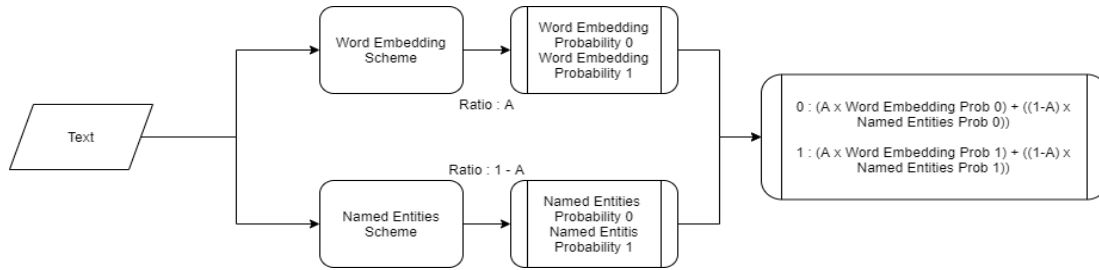


Figure 2. Hybrid scheme concept

4. RESULTS AND DISCUSSION

4.1. Data collection

Entity tagging is the initial step before preprocessing. This enables the creation of diverse texts containing things that have significance but are deleted during preprocessing and/or are poorly comprehended by computers. Data from crawl results, as shown in Table 1, will be cleaned through many steps of the preprocessing model. The lower () method from Python's string module is used to fold the cases. Using the Python string package, superfluous characters such as emoticons, website URLs, punctuation marks, double spaces, and newlines are removed. Because the natural language tool kit (NLTK) library does not currently support Indonesian, the stemming process in Indonesian is carried out using the sastrawi library, which has shown to be fairly excellent at handling the Indonesian language stemming process. The NLTK library is used in the tokenizing process to divide sentences into lists with a space character separator. The NLTK and sastrawi libraries are used in the stopword elimination procedure. The stopword removal procedure will be strengthened by using two libraries, which will compensate for each other's inadequacies.

Table 1. Example of crawling process result

Column	Example
Created_At	Thu Feb 22 18:09:57 +0000 2021
Id	1397978397960065024
Full_Text	Dua Truk Adu Banteng Di Pati Bermula Saat Hino Coba Salip Motor, Begini Kronologinya.\N\nselengkapnya Klik Tautan Berikut Ini. \N#Pati #Kronologi #Kecelakaan #Truk \N\Nhttps://T.Co/Vk1prhhdzp

4.2. Named-entities tagging

Labeling named entities for the terms in the dataset completes this phase. In this study [22], the specified entity relates to various name-entities connected with traffic. This phase is completed by labeling named entities for terms in the data collection. In this study [22], the specified entity corresponds to various name entities that are associated with traffic. Table 2 lists the named entity categories that have been defined and have a strong relationship with the accident data. When tagging, the outcome of this group is utilized to create a named entity label group. The researcher created the labeling application using the Laravel framework [23] and the PostgreSQL database. Entity tagging is done on data that has been acquired in a certain length of time. Figure 3 and Figure 4 depict the outcomes of the tagging procedure.

Table 1. List of named-entities annotated

Entity	Name	Example
DAT	Date	September, 2019, Besok
LOC	Location	Rawamangun, Jakarta, China, Cipali, Semarang
ORG	Organization	Lion Air, BUMN, Polri, Kemenhub
TIM	Time	15.24, Pagi, Malam
VEH	Vehicle	Avanza, Innova, Boeing, Mobil, Bus

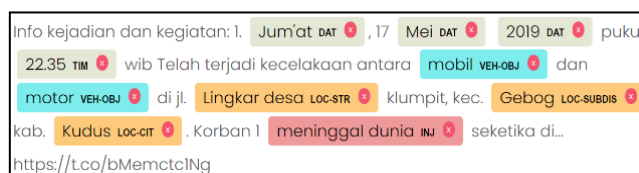


Figure 1. Entity tagging process

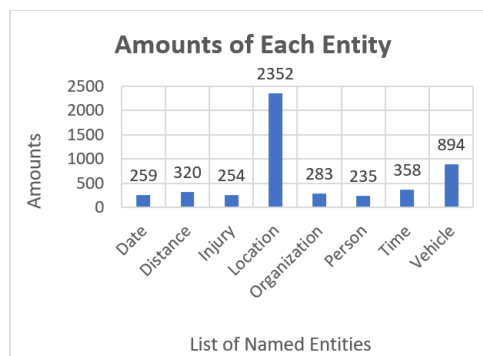


Figure 4. Amounts of each entity

4.3. Data pre-processing

Data processing is one of the most important aspects of the data analysis process, and it frequently necessitates more work and time [24]. In this phase, we'll tag named entities. This enables the creation of diverse texts containing things that have significance but are deleted during preprocessing and/or are poorly comprehended by computers. Data will be cleaned through many steps of the preprocessing model. The lower() method from Python's String module is used to fold the cases. Using the Python String package, superfluous characters such as emoticons, website URLs, punctuation marks, double spaces, and newlines are removed. The stemming process is an important pre-processing phase that, depending on the language employed, might be considered a tough step to complete. The amount of morphological complexity of a language can impact stemming outcomes [25]. Because the NLTK library [26], which is used for the stemming process, does not currently support Indonesian, the stemming process in Indonesian is carried out using the Sastrawi library [27], which has proved to be fairly competent in handling the Indonesian language stemming process. The NLTK library is used in the tokenizing process to divide sentences into lists with a space character separator. The NLTK and Sastrawi libraries are used in the stopword elimination procedure. The stopword removal procedure will be strengthened by using two libraries, which will compensate for each other's inadequacies. The number of entities calculated in each report is used as a parameter in the named entities and hybrid scheme. The FastText word embedding procedure is carried out with the use of pre-trained Indonesian language models, which may be found at FastText's website [28]. Emoji removal, punctuation removal, case folding, stemming, stopword, tokenization, and representation of FastText words are all steps of data pre-processing that are executed using the Word2Vec model. The number of entities calculated in each report is used as a parameter in the NE and combination modeling techniques. Table 3 shows the outcomes of the pre-processing.

Table 3. The results of the pre-processing

Processed Text	Vectorized Text	Entity Count	Is Accident
[bruk, kecelakaan, maut, libatkan, 2, mobil, 1...	[0.117599905, 0.17269996, 0.9895, 0.51780003, ...	[2, 0, 0, 0, 1, 3]	1
[13, 14terjadi, kecelakaan, beruntun, jl, mayj...	[-0.06250001, 0.3759, -0.045600012, -0.0926, 0...	[1, 0, 1, 0, 0, 0]	1
[kecelakaan, siang, jl, lahor, batu, kejadian,...	[0.20839998, -0.07299999, -0.5571, 0.7532, 0.5...	[1, 0, 1, 0, 0, 1]	1
[kecelakaan, jl, raya, serang, pandeglang, tep...	[-0.58949995, 0.09609996, 0.2997, 0.26459998, ...	[2, 1, 2, 0, 0, 0]	1
[jujutsufess, childhood, friend, merangkap, cr...	[1.8680998, -1.1477001, 0.44889998, 1.4204, -0...	[0, 0, 0, 0, 0, 0]	0

4.4. Training classification using the SVM algorithm

The three techniques mentioned in the preceding section are used to classify the data:

- Word embedding: The FastText Word Embedding model is used to provide the position value for each text when modeling using word representation.
- Named entities: The quantity of each named entity in a text is used to identify the mix of entities in a text when modeling with entity tagging.
- Hybrid: Combination is achieved by combining the two models mentioned above, which then predicts a text by comparing each model's contribution.

The K-fold cross validation technique is used to validate the training outcomes. Cross validation is a technique that provides a systematic way for assessing model efficacy and comparing models to one another. This technique assumes that the model was trained on a separate dataset from the one that was used for testing. The model finds rules in one dataset and then values them in a another dataset. Model accuracy may

be objectively verified using the validation dataset, which provides information on genuine classification results [29]. The dataset for this procedure will be derived via data validation. This technique divides the dataset into ten sections and changes locations ten times as a 90% training fold and 10% validation fold.

In this scheme, we examined the contribution ratios of each scheme, which ranged from 0.2 to 0.8. As a consequence, the best accuracy comparison was achieved when the word embedding scheme and the named entities scheme were combined at 0.85 vs 0.15. As a result of this comparison, the named entities scheme may give probabilities as a supplement to the hybrid scheme while maintaining a balanced contribution ratio value. The combination strategy is shown in Table 4, with SVM surpassing the other two by a score of 90.27%. This demonstrates that using named entities in the traffic accident report data categorization process as a supporting scheme for word embedding has resulted in a 2.70% increase in capabilities. The hybrid scheme has a cross-validation score of 81.98%. This demonstrates that the hybrid approach works effectively with fresh data.

Table 2. Results of schema classification

Schema	Accuracy Score	Cross Validation
Word Embedding	0.875676	0.808069
Named Entities	0.810811	0.794828
Hybrid	0.902703	0.811595

5. CONCLUSION AND FUTURE WORK

The dataset includes of data in its original state, data labeling results for defined entities, pretreatment processing results, and word embedding representation results. An evaluation of the performance of each scheme is carried out with a model accuracy score based on the suggested modeling scheme to examine the influence of named entities on the categorization of traffic accident data. When using a hybrid strategy with the SVM model, the best accuracy results are obtained at 90.27%. This approach outperforms the categorization method based on traditional word embedding, which scored 87.57% in this study's comparison. It's likely that the named entity scheme provides explanation to the comprehension of sentences that aren't effectively represented by word embedding, allowing the result to improve. However, using the number of occurrences of named entities as an only input for text categorization produced poor results, with the lowest score of 81.08% when compared to alternative techniques.

This dataset can be acquired and used in the future for research. Labeling the data for training is required to improve machine learning with a broader data range. Additional machine learning approaches, such as deep learning, can be enabled by incorporating sufficient training data. It's also possible to broaden the labeling options for Named Entities. Because the proposed hybrid method largely depends on data from named entity labels, accurate labeling of named entities is required to offer good sentence interpretation. It is believed that the computer would be able to comprehend the meaning of the word in its context in greater depth, while remaining a single entity. Collecting data from sources other than Twitter, on the other hand, is advised in order to create bigger and more varied databases.





REFERENCES

- [1] R. Ishrat, "Global Status Report on Road Safety 2018: Summary," *World Health Organization*, no. 1, p. 20, 2018, Accessed: Jan. 19, 2022. [Online]. Available: <http://apps.who.int/bookorders>.
- [2] R. Sujay, J. Pujari, V. S. Bhat, and A. Dixit, "Timeline Analysis of Twitter User," *Procedia Computer Science*, vol. 132, pp. 157–166, 2018, doi: 10.1016/j.procs.2018.05.179.
- [3] D. A. Saputro and A. S. Girsang, "Classification of traffic accident information using machine learning from social media," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 3, pp. 630–637, Mar. 2020, doi: 10.30534/ijeter/2020/04832020.
- [4] A. V. Deursen, A. Mesbah, and A. Nederlof, "Crawl-based analysis of web applications: Prospects and challenges," *Science of Computer Programming*, vol. 97, no. P1, pp. 173–180, Jan. 2015, doi: 10.1016/j.scico.2014.09.005.
- [5] D. Antonakaki, P. Fragopoulou, and S. Ioannidis, "A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks," *Expert Systems with Applications*, vol. 164, p. 114006, Feb. 2021, doi: 10.1016/j.eswa.2020.114006.
- [6] "Twitter's 10 Year Struggle with Developer Relations | Nordic APIs |." <https://nordicapis.com/twitter-10-year-struggle-with-developer-relations/> (accessed Apr. 27, 2021).
- [7] "Delivering a consistent Twitter experience." https://blog.twitter.com/developer/en_us/a/2012/delivering-consistent-twitter-experience (accessed Apr. 27, 2021).
- [8] "The Infrastructure Behind Twitter: Scale." https://blog.twitter.com/engineering/en_us/topics/infrastructure/2017/the-infrastructure-behind-twitter-scale (accessed Apr. 28, 2021).
- [9] K. Takeuchi and N. Collier, "Bio-medical entity extraction using Support Vector Machines," in *Proceedings of the {ACL} 2003 workshop on Natural language processing in biomedicine -*, 2003, pp. 57–64, doi: 10.3115/1118958.1118966.
- [10] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.





- [11] A. Goyal, V. Gupta, and M. Kumar, "Recent Named Entity Recognition and Classification techniques: A systematic review," *Computer Science Review*, vol. 29, pp. 21–43, Aug. 2018, doi: 10.1016/j.cosrev.2018.06.001.
- [12] M. Paşca, D. Lin, J. Bigham, A. Lifchits, and A. Jain, "Organizing and searching the World Wide Web of facts - Step one: The que-million fact extraction challenge," *Proceedings of the National Conference on Artificial Intelligence*, vol. 2, 2006, pp. 1400–1405, Accessed: Jan. 19, 2022. [Online]. Available: <https://research.google/pubs/pub69/>.
- [13] J.-H. Kim, I.-H. Kang, and K.-S. Choi, "Unsupervised named entity classification models and their ensembles," in *Proceedings of the 19th international conference on Computational linguistics -*, 2002, pp. 1–7, doi: 10.3115/1072228.1072316.
- [14] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel, "Nymble: A high-performance learning name-finder," in *5th Conference on Applied Natural Language Processing, ANLP 1997 - Proceedings*, 1997, pp. 194–201, doi: 10.3115/974557.974586.
- [15] A. Sultan, A.-H. Ameen, M. Farea, O. Fuad, and T. Bagash, "A Biomedical Named Entity Recognition Using Machine Learning Classifiers and Rich Feature Set," *IJCSNS International Journal of Computer Science and Network Security*, vol. 17, no. 1, p. 170, 2017.
- [16] J. Li, A. Sun, J. Han, and C. Li, "A Survey on Deep Learning for Named Entity Recognition," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2020, doi: 10.1109/tkde.2020.2981314.
- [17] J.-H. Kim and P. Woodland, "A rule-based named entity recognition system for speech input," 2000, pp. 528–531.
- [18] "Twitter: monthly active users worldwide | Statista," <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/> (accessed Jul. 03, 2020).
- [19] A. Gal-Tzur, S. M. Grant-Muller, T. Kuflik, E. Minkov, S. Nocera, and I. Shoor, "The potential of social media in delivering transport policy goals," *Transport Policy*, vol. 32, pp. 115–123, Mar. 2014, doi: 10.1016/j.tranpol.2014.01.007.
- [20] Y. Gui, Z. Gao, R. Li, and X. Yang, "Hierarchical text classification for news articles based-on named entities," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7713 LNAI, Springer Berlin Heidelberg, 2012, pp. 318–329.
- [21] Y. Tan, "Applications," in *Gpu-Based Parallel Implementation of Swarm Intelligence Algorithms*, Elsevier, 2016, pp. 167–177.
- [22] M. Schiersch, V. Mironova, M. Schmitt, P. Thomas, A. Gabryszak, and L. Hennig, "A German corpus for fine-grained named entity recognition and relation extraction of traffic and industry events," *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, Apr. 2019, pp. 4437–4444, Accessed: Jan. 19, 2022. [Online]. Available: <https://arxiv.org/abs/2004.03283v1>
- [23] R. Y. He, "Design and Implementation of Web Based on Laravel Framework," in *Proceedings of the 2014 International Conference on Computer Science and Electronic Technology*, vol. 6, 2015, doi: 10.2991/iccset-14.2015.66.
- [24] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, and F. Herrera, "A survey on data preprocessing for data stream mining: Current status and future directions," *Neurocomputing*, vol. 239, pp. 39–57, May 2017, doi: 10.1016/j.neucom.2017.01.078.
- [25] M. Naili, A. H. Chaibi, and H. H. Ben Ghezala, "Comparative study of Arabic stemming algorithms for topic identification," *Procedia Computer Science*, vol. 159, pp. 794–802, 2019, doi: 10.1016/j.procs.2019.09.238.
- [26] V. N. Gudivada and K. Arbabifard, "Open-Source Libraries, Application Frameworks, and Workflow Systems for NLP," in *Handbook of Statistics*, vol. 38, Elsevier, 2018, pp. 31–50.
- [27] "sastrawi/sastrawi: High quality stemmer library for Indonesian Language (Bahasa)," <https://github.com/sastrawi/sastrawi> (accessed Apr. 28, 2021).
- [28] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, Dec. 2017, doi: 10.1162/tacl_a_00051.
- [29] M. Rafalo, "Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis," *ICT Express*, May 2021, doi: 10.1016/j.ict.2021.05.001.

BIOGRAPHIES OF AUTHORS



Anugrah Dwiatmaja Putra     is currently a software engineer and project manager at a national company. He earned his M. Kom. at Bisa Nusantara University, Department of Informatics Engineering, Jakarta Indonesia, in 2021, and completed his undergraduate education from the Department of Information Systems, Sepuluh Nopember Institute of Technology, Surabaya Indonesia, in 2018. He was a Full Stack Developer at Riliv.co, Surabaya, in 2016–2018 and also worked as a web developer in various projects in 2015–2019. He can be contacted at email: anugrahdputra@gmail.com.



Abba Suganda Girsang     is currently lecturer at master information technology at Bina Nusantara University Jakarta. He obtained Ph.D. degree in the Institute of Computer and Communication Engineering, Department of Electrical Engineering and National Cheng Kung University, Tainan, Taiwan, in 2014. He graduated bachelor from the Department of Electrical Engineering, Gadjah Mada University (UGM), Yogyakarta Indonesia, in 2000. He then continued his masters degree in the Department of Computer Science in the same university in 2006–2008. He was a staff consultant programmer in Bethesda Hospital, Yogyakarta, in 2001 and also worked as a web developer in 2002–2003. He then joined the faculty of Department of Informatics Engineering in Janabadra University as a lecturer in 2003–2015. He also taught some subjects at some universities in 2006–2008. His research interests include swarm intelligence, combinatorial optimization, and decision support system. He can be contacted at email: agirsang@binus.edu.