Bertopic And NER Stop Words For Topic Modeling On Agricultural Instructional Sentences

Trisna Gelar and Aprianti Nanda Sari
Department of Computer and Informatics Engineering
Politeknik Negeri Bandung, POLBAN
Bandung, Indonesia
trisna.gelar@polban.ac.id, aprianti.nanda@polban.ac.id

Abstract. A drawback of topic modeling is the lack of consistent sentence frequency within each topic. The outcome of this event manifests as varying levels of topic coherence and topic diversity. One potential approach to addressing this issue involves the modification of stop words, which refers to the removal of unneeded or excessively utilized terms. In the context of specialist areas like health, law, and agriculture, the identification of stop words can be achieved through the utilization of Name Entity Recognition (NER). This procedure involves preprocessing the data before subjecting it to topic modeling. Furthermore, it is possible to investigate the utilization of several topic modeling elements in conjunction with BERTopic to enhance the efficacy of the generated topics. The most effective configuration for the BERTopic pipeline consists of employing Sentence Embedding for text representation, UMAP Dimensionality Reduction for feature reduction, HDBScan Clustering for grouping similar documents, and utilizing a combination of Named Entity Recognition (NER) for removing stop words and C-TF-IDF for topic representation. This has resulted in the highest level of topic diversity performance for JADI and PUW by 0,982 and 0,990. The method generated the minimum number of outliers. However, there has been a decrease in the effectiveness of topic coherence.

Keywords.BERTopic; NER; Stop Words; Topic Modeling;

1. Introduction

Topic modeling is an attempt to cluster sentences or documents based on probabilistic calculations of building words (word embedding) into a related subject. Bag of Words (BoW) is one of the most frequently employed word embedding representations in Topic Modeling algorithms like Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-negative Matrix Factorization (NMF). These three methods are efficient and practical, but the resulting topic model fails to represent the syntactic and semantic relationships between words. Due to an incoherent BoW word representation will result in an inaccurate topic representation[1].

Transformer-based document-embedding representation is an alternative method. In contrast to BoW, document-embedding approaches such as Bidirectional Encoder

M. U. H. Al Rasyid and M. R. Mufid (eds.), *Proceedings of the International Conference on Applied Science and Technology on Engineering Science 2023 (iCAST-ES 2023)*, Advances in Engineering Research 230,
https://doi.org/10.2991/978-94-6463-364-1_14

Representations from Transformers (BERT)[2] are able to produce coherent word representations. BERTopic[3] is a technique for topic modeling that adopts BERT document-embedding representations as its word-building method. BERTopic can be utilized in short texts[4] such as Twitter tweets, e-commerce review comments, and instructional texts, as well as documents with complex structures.

Several researchers have conducted experiments and implemented BERTopic in various problem areas to model the topic of short texts. Mahfudiyah and Alamsyah[5] conducted an analysis of consumers' perceptions of Gojek services; the data came from Twitter tweets in Indonesian; the process consisted of two stages: sentiment analysis of IndoBERT tweets that were positive, negative, and neutral; and topic modeling with BERTopic for every sentiment. The classification of Indonesian hoax news by Hutama and Suhartono[6] involved two primary stages: feature extraction using a pretrained multilingual model (XML-R, mBERT) and topic distribution using BERTopic. Because classification was the objective, the researcher chose the five and ten most frequent words from each of the 24 topics to improve the results of hoax classification. However, the performance of the generated topics and the configuration of the BERTopic implementation were not included in the two studies.

The topic modeling performed by Metayasa and Darmawan[7] on English hotel review data from TripAdvisor generated 78 topics with a relatively low performance value (topic coherence of 0.072 and topic diversity of 0.496). While Priyatna [8] performed topic modeling on English video game review data, using only Not Recommended reviews to provide game developers with constructive feedback on how to enhance their games. In addition, the Guided topic modeling procedure is chosen, resulting in ten topics with a topic coherence value of 0.524 percent.

The issue with topic modeling is that the frequency of sentences within each topic is inconsistent. This results in variable topic coherence and topic diversity performance values. For example, a model may have a high topic coherence value, but after analysis, it may contain many sentences that are outliers and are therefore not representative. One solution to this problem is to modify stop words[9], i.e., to eliminate unnecessary or overused words. In documents with specialized domains such as health, law, and agriculture, stop words can be identified using Name Entity Recognition (NER) by preprocessing the data prior to topic modeling processing[10] In addition, the use of topic modeling components such as document-embedding configuration, dimension reduction method, clustering method, use of tokenization (number of n-grams), and the c-TF-IDF weighting scheme can be explored with BERTopic to improve the performance of the resulting topics.

In this study, agricultural instructional texts became a case study for the development of topic modeling; urban agriculture documents on horticultural topics were curated alongside their NER entities[11]. The vocabulary used in agricultural instructional texts includes biological and chemical elements, agricultural instruments, diseases, agricultural products, and agricultural methods, among others. The entity will be combined as a stop word with BERTopic to produce an interpretable topic model.

2. Related works

2.1 Topic Modeling

BoW-based topic modeling has two disadvantages, namely that the final model is unable to recognize syntactic and semantic relationships between words in a document[1]. Input to the BoW that lacks coherence will result in inaccurate topic representation[3]. For instance, a document-embedding approach to generating coherent models can reduce BoW issues. Transformer-based representation languages, such as Bidirectional Encoder Representations from Transformers (BERT), have been evaluated for their ability to construct language representations from extremely large text corpora in a variety of scenarios.

2.2 The Role of NER in Topic Modeling

Katsiaryna Krasnashchok, et al. [10] used entities in specific domains to assist present topic models by giving entities greater weight than words containing NER. The use of NER in modeling topics has not received much attention. In their research, the LDA method was utilized, with experimental results indicating that the inclusion of more named entities in topic descriptions increases topic coherence and topic diversity.

In another study, Yau et al. [12] used NER as part of word preprocessing prior to performing transformer-based document embedding to derive topic modeling with BERTopic for the health problem domain. Christensen et al. [13] implement BERTopic and NER for the same domain to determine public adoption of herd immunity cases through COVID-19 vaccination; however, NER is used for news interpretation or the post-processing stage to determine the subject and object of the news.For specialized domains, BERTopic combined with Name Entity Recognition techniques will have a positive effect on topic interpretation, as the corpus containing specialized terms is only partially understood, and NER can close the knowledge gap.

2.3 Stopwords on Topic Modeling

Schofield, Magnusson, and Mimno[9] investigated the relationship between stop words and Topic modeling. There are three hypotheses for this problem: stop words can hinder topic inference, stop words have no influence on topic inference. By eliminating stop words, it is assumed that the model will learn from high-quality data to generate relevant topics and keywords.According to the results of the research, removing stop words can enhance the quality of the topic, but the effect is superficial if the removed word is not a frequent one in the document. NER entities can be used as stop words as part of an endeavor to identify high-frequency words in specific domains. However, it must be determined which entities can enhance topic inference quality. In addition, removing stop words during preprocessing and postprocessing has the same effect on topic modeling from a technical standpoint.

2.4 Topic Modeling Performance

Topic coherence is a performance indicator that measures the quality of the topic representation. The topics will be cogent if the words that appear are consistently found in the same document or sentence, or if there is a high degree of informational connectivity. The higher the score (0 to 1), the simpler it is to comprehend the topic. Roder et al. [14] proposed a framework for measuring topic coherence comprising four stages: segmentation, probability calculation, confirmation measure, and aggregation.

The genism topic modeling library offers topic coherence measurements based on the coherence model framework, specifically Cv, CUmas, and CNPMI[15].

Topic Diversity is a performance metric for evaluating the quality of the model being studied. In proportion to how different or unique a topic is, its proportional representation grows. There are two measures of topic diversity: average Pairwise Jaccard Distance and proportion of unique terms.

Pairwise Jaccard Distance (JADI) is a topic diversity measure based on the Jaccard coefficient [16]. The proportion of unique words (PUW) is a measure of topic diversity based on the proportion of unique words in the topic representation. If the PUW value is close to zero, the topic is redundant; otherwise, it varies significantly[17].

Health[18], tourism[19] and other domains have been modeled using BERTopic-based topic modeling. BERTopic is a topic modeling technique that seeks to generate topic clusters that are easily interpretable. Figure 1 depicts the BERTopic procedure, which involves reducing the dimensions of BERT with the UMAP method and generating clusters with HDBSAN. The model then determines the top five words or keywords based on class-based TF-IDF (c-TF-IDF) by identifying the unique and significant words within each cluster.

BERTopic utilizes an embedding-based language model that can comprehend the context of each word. The words “kali” in the sentences " Buah mangga jatuh di pinggir **kali**" and " Pemupukan NPK pada mangga dilakukan dua **kali** sehari " can be distinguished by BERT because each word has a distinct embedding value based on the context of the sentence. For each model document or sentence, the subject. By utilizing pretrained models, the Sentence-BERT Framework (SBERT)[20] converts sentences or paragraphs into dense vector representations.

2.5 BERTopic

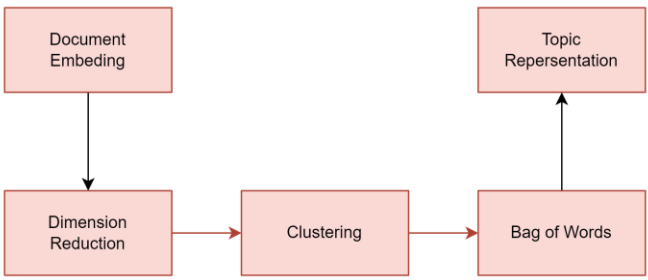

Fig. 1. BERTopic Procedure

The procedure of dimensionality reduction is performed to save space in the embedding document so that it can be searched efficiently and precisely. There are three applicable algorithms for BERTopic: UMAP, PCA, and SVD. Uniform Manifold Approximation and Projection (UMAP) is able to preserve local and global features in high-dimensional data after compression into low-dimensional data and provides greater scalability than PCA or SVD[21]. There are two primary hyperparameters for configuring UMAP: the number of nearest neighbors, which controls the proportion

between global and local features, and the number of nearest neighbors. While the minimum distance controls how closely the algorithm collects data, it regulates how close the distance is.

On reduced document embedding, the clustering operation is conducted. Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), K-Means Clustering, and Agglomerative Clustering are the three algorithms that can be applied to BERTopic. HDBSCAN excels at locating clusters of varying densities and sizes, even in the presence of noise or outliers. HDBSCAN enables noise data to exist in a class of its own, preventing noise from interfering with the algorithm's determination of the topic, since prominent noise lacks descriptive capabilities. The most important hyperparameter in HDBSCAN is the minimum number of clusters, which determines the minimum size that must be deemed a cluster by the algorithm[22].

C-TF-IDF is a variation of the class-based Term Frequency Inverse Document Frequency (TF-IDF) utilized by BERTopic. The TF-IDF calculation is determined by comparing the relative frequency of each word in a document to the inverse proportion of the document's words. The calculation represents a word's importance in the document. To search for significant words in document clusters in a similar manner, all documents belonging to a particular cluster will be combined to form a cluster class[3].

3. Material and Method

3.1 Dataset Agricultural Instructional Sentences

Documents of Agriculture Instructional Text The Horticulture (fruits, floriculture, vegetables, and medicinal plants) served as a source of texts for language model development. Each volume is in PDF format, requiring additional processing to extract the specific text of agricultural terms. In addition, the author performs manual preprocessing, such as creating bookmarks and separating less important sections of the book, such as covers, prefaces, tables of contents, tables that are difficult to extract directly by computers, and bibliographies. The data for Agricultural Instructional Text and NER Entity is derived from research[23][11]. In total there are 960 agricultural instructional sentences and 10 NER Entity.

TABLE 1. AGRICULTURAL NER

Entity	Description	Frequency
CROP	Names of Fruits and Plants	322
CHEMICAL	Chemical Elements, Fertilizers, and Pepticides	282
QUANTITY	Weight measures, such as liters, tons and grams.	154
DISTANCE	Distance Measures, such as cm, m, and km	114
DISEASES	Diseases of fruits and plants.	33
PESTS	Pests that cause disease in fruits and plants	194

LOCATION	The name of a city, state, area or place.	214
CARDINAL	The amount of an object	846
VARIETIES	Varieties of plants or fruits.	101
PERIOD	Frequency or cycles of a method / process.	178

3.2 Research Method

In this study, quantitative methods are employed, and the performance of topic modeling using BERTopic and NER Stopwords will be empirically evaluated. Figure 2 is an overview of the stages of the research, which include the initiation, implementation, and evaluation phases.

- During the Initiation phase, the Agricultural Instructional Text data will be filtered for the specific context of fruit cultivation (Orange, Passion Fruit, Durian, Dragon Fruit, and Mango). In addition, each document will be divided into sentences or paragraphs with complete context. In addition, during the Preprocessing phase, each sentence will undergo a data cleansing procedure, such as cascading, symbol removal, stop word removal, and stemming. Using NER Agriculture[11], data enhancement is performed so that each sentence contains stop words or high-frequency entities.
- In the phase of implementation, topic modeling will be performed using the BERTopic algorithm, which is comprised of four primary stages: Document Embedding, Dimensionality Reduction, Clustering, and Topic Representation. Specifically for Indonesian Document Embedding, the use of pretrained-model options is still limited. The utilized multilingual pretrained models are listed in Table II. For topic modeling, the three pretrained models will be investigated. In the process of Dimensionality Reduction and Clustering, a combination of PCA and UMAP with HDBSCAN, K-Means, and Agglomerative Clustering will be implemented and evaluated. C-TF-IDF will be used specifically for topic representation in this study, with the option of reducing words that frequently occur on cluster topics. The combinations of BERTopic pipeline are listed in Table III.
- In the Evaluation phase, the topic modeling outcomes for each combination of BERTopic components will be empirically measured. The measurement of performance centers on topic coherence and topic diversity. Plan for evaluating topic modeling is shown in Table III. The evaluation results will be analyzed to determine the optimal combination of components or to identify extreme results from the topic modeling. Moreover, for each discovery, discussion and discussion were conducted.

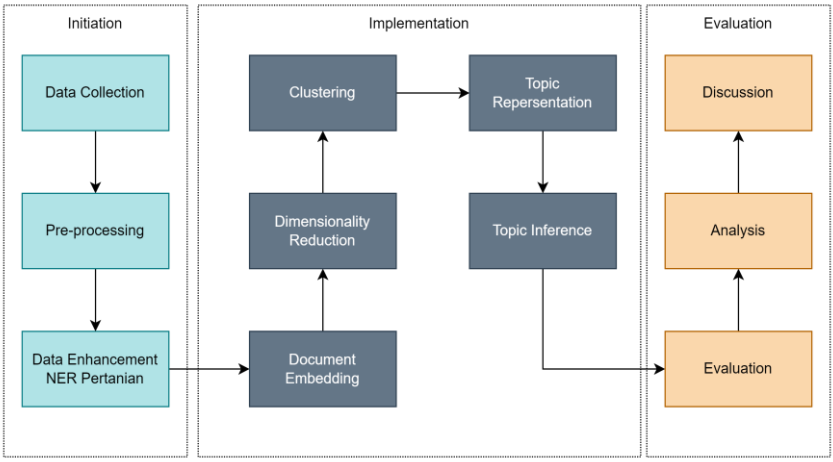

Fig. 2. Research Method

TABLE 2. PRETAINED LANGUAGE MODEL

Pretained Model	Description
distiluse-base-multilingual-cased-v2 (default)	multi-language more than 50 language, implemented with Sentence-BERT.
paraphrase-multilingual-MiniLM-L12-v2 (Sentence Embeding)	multi-language more than 50 language, implemented with Sentence-BERT.
indobenchmark/indobert-base-p1 (IndoBERT)	Indonesia language, implemented with flair embedding[24]

TABLE 3. BERTOPIC PIPELINE COMBINATION FOR EXPERIMENTS

Data	BERTopic Pipeline		
	Clustering	Dimensionality Reduction	Topic Representation
Without NER	UMAP	HDBSCAN	C-TF-IDF
With NER		K-Means	
		Agglomerative Clustering	

4. Result And Discussion

4.1 Result

The results acquired from all the conducted experimental scenarios. There are three experimental categories determined by the combination of BERTopic pipelines and their impact on topic modeling performance (topic coherence and topic diversity). First, the document embedding pipeline utilizes multilingual or Indonesian language pretrained

models (Table IV). Second, The Clustering algorithm pipeline employs three methods, namely Agglomerative Clustering, K-Means Clustering, or HDBSCAN (Table V). Third, the C-TF-IDF topic representation pipeline with or without NER stop words (Table VI). The use of sentence embedding as the document embedding pipeline in has the highest performance value for topic modeling compared to other pretrained models. Which is TC, JADI and PUW by 0.73, 0.911, and 0.964, respectively. This model will become the standard for the next combination pipeline.

TABLE 4. BERTOPIC MODEL WITH DIFFERENT DOCUMENT EMBEDDING PIPELINE

Document Embedding	N Topic Reduction	Outlier	Performance		
			TC	JADI	PUW
Default	14	178	0,691	0,865	0,972
Sentence Embedding	15	353	0,731	0,911	0,964
IndoBERT	14	450	0,700	0,865	0,966

The use of HDBScan as the Clustering pipeline in has resulted in the highest topic modeling performance value when compared to other methods. Which is TC, JADI, and PUW with respective values of 0.698, 0.893, and 0.96. This model will serve as the benchmark for the subsequent combination pipeline. It is also important to note that the number of outliers was lower than that of other models. Experimentation with the Topic Representations pipeline utilizing C-TF-IDF and combined with NER resulted in a 0.982- and 0.990-performance enhancement for JADI and PUW, respectively. In addition, the algorithm produced the fewest outliers compared to all other experiments. Nevertheless, the efficacy of topic coherence has declined.

TABLE 5. BERTOPIC MODEL WITH DIFFERENT CLUSTRING PIPELINE

Pipeline Combination	N Topic Reduction	Outlier	Performance		
			TC	JADI	PUW
Agglomerative Clustering-UMAP	15	448	0,665	0,857	0,952
K-Means UMAP	15	569	0,658	0,875	0,959
HDBScan UMAP	15	340	0,698	0,893	0,968

TABLE 6. BERTOPIC MODEL WITH C-TF-IDF AND NER AS STOP WORDS

Pipeline Combination	N Topic Reduction	Outlier	Performance		
			TC	JADI	PUW
C-TF-IDF without NER	15	304	0,639	0,946	0,986

C-TF-IDF with NER	15	236	0,624	0,982	0,990

4.2 Discussion

According to the results of the experiments, the usage of IndoBert had no significant effect on the results of modeling agricultural instructional text topics compared to multilingual sentence embedding. This is since the themes in IndoBert are still universal and not tailored to agricultural problems. The HDBScan approach in pipeline clustering can greatly reduce outliers when compared to K-Means and Agglomerative Clustering methods and is superior for all measures of TC, SO, and PUW performance. The optimal combination of the BERTopic pipeline is Sentence Embedding, UMAP Dimensionality Reduction, HDBScan Clustering, and the combination of NER Stop words and C-TF-IDF topic representation. This has produced the best JADI and PUW topic diversity performance. Despite having the lowest topic coherence, the outlier is reduced to 236 sentences. The outlier was generated by UMAP Dimensionality Reduction, and the larger number of outliers generated appears to enhance the TC value because these outlier sentences were not included in the creation of topic modeling word structures. Figure 3 and Figure 4 depict hierarchical clusters of the top-10 topic combination words without and with NER as stop words.

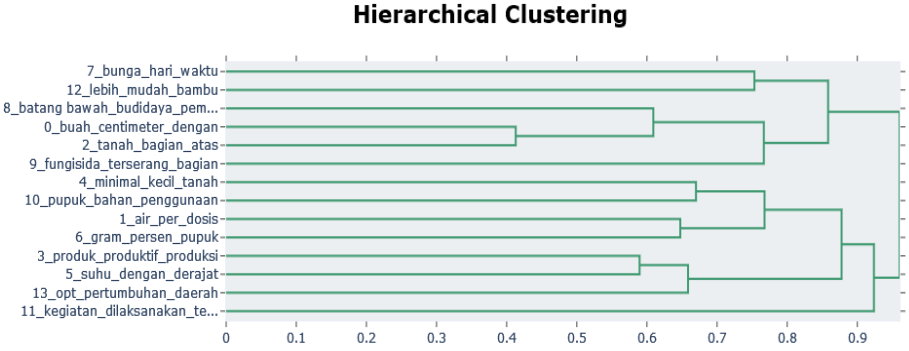

Fig. 3. Hierarchial Clustering Topic Models C-TF-IDF without NER

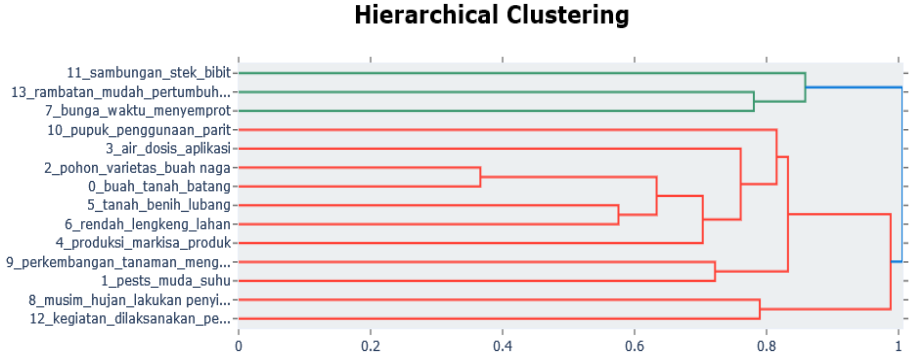

Fig. 4. Hierarchial Clustering Topic Models C-TF-IDF with NER

5. Conclusion

Various configurations of BERTopic pipelines have been employed to develop topic modeling for agricultural instruction. The optimal pipeline for document embedding and clustering employs sentence embedding and HDBScan, followed by UMAP Dimensionality Reduction, yielding the highest topic modeling performance scores across the three-performance metrics. Furthermore, agricultural instructional sentences that were preprocessed with stop words using NER and combined with C-TF-IDF topic representations had an effect on the performance of topic diversity (JADI and PUW) and reduced the number of outliers produced through UMAP, however had no influence on the performance of topic coherence. For future research to enhance topic coherence performance, a mechanism for selecting optimal NER entities for stop words, utilizing a swarm algorithm would be helpful.

ACKNOWLEDGMENT

The Research Program is supported by DIPA funds from the Bandung State Polytechnic in accordance with a letter of agreement the PM number B/98.48/PL1.R7/PG.00.03/2023. Thank you to everyone who contributed to the success of this endeavor.

REFERENCES

[1] F. Bianchi, S. Terragni, and D. Hovy.: "Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2021, pp. 759–766, doi: 10.18653/v1.acl-short.96(2021)..

[2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova.: "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, no. Mlm, pp. 4171–4186, (2019).

[3] M. Grootendorst.: "BERTopic: Neural topic modeling with a class-based TF-IDF procedure,", [Online]. Available: http://arxiv.org/abs/2203.05794(2022).

[4] M. de Groot, M. Aliannejadi, and M. R. Haas.: "Experiments on Generalizability of BERTopic on Multi-Domain Short Text," pp. 1–3, [Online]. Available: http://arxiv.org/abs/2212.08459 (2022).

[5] N. Mahfudiyah and A. Alamsyah.: "Analisis Persepsi Konsumen Terhadap Kualitas Layanan Gojek Menggunakan Sentiment Analysis Dan Topic Modeling Berdasarkan Deep Learning IndoBERT," *e-Proceeding of Management*, vol. 9, no. 4, pp. 1812–1817, (2022).

[6] L. B. Hutama and D. Suhartono.: "Indonesian Hoax News Classification with Multilingual Transformer Model and BERTopic," *Informatica (Slovenia)*, vol. 46, no. 8, pp. 81–90, doi: 10.31449/inf.v46i8.4336(2022).

[7] I. K. Tryana, I. D. Made, and B. Atmaja.: "Pemodelan Topik Pada Ulasan Hotel Menggunakan Metode BERTopic Dengan Prosedur c-TF-IDF," vol. 1, no. November, pp. 307–316, (2022).

[8] G. G. Priyatna.: "Pemodelan Topik Terkait Ulasan Video Game Dengan Genre Battle Royale

Menggunakan Metode BERTopic dengan Fitur Guided Topic Modeling," Universitas Islam Negeri Syarif Hidayatullah, (2023).

[9] A. Schofield, M. Magnusson, and D. Mimno.: "Pulling out the stops: Rethinking stopword removal for topic models," *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, vol. 2, no. January, pp. 432–436, doi: 10.18653/v1/e17-2069(2017).

[10] K. Krasnashchok and S. Jouili.: "Improving topic quality by promoting named entities in topic modeling," *ACL 2017- 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 2, pp. 247–253, doi: 10.18653/v1/p18-2040(2018).

[11] T. Gelar, A. Nanda, and A. Bakhrun.: "Serverless Named Entity Recognition untuk Teks Instruksional Pertanian Kota," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, no. 3, pp. 597–606, doi: 10.28932/jutisi.v8i3.5447(2022).

[12] Q. X. Ng, C. E. Yau, Y. L. Lim, L. K. T. Wong, and T. M. Liew.: "Public sentiment on the global outbreak of monkeypox: an unsupervised machine learning analysis of 352,182 twitter posts," *Public Health*, vol. 213, pp. 1–4, Dec. 2022, doi: 10.1016/j.puhe.2022.09.008(2022).

[13] B. Christensen *et al.:* "Quantifying Changes in Vaccine Coverage in Mainstream Media as a Result of the COVID-19 Outbreak: Text Mining Study," *JMIR Infodemiology*, vol. 2, no. 2, doi: 10.2196/35121(2022).

[14] M. Röder, A. Both, and A. Hinneburg.: "Exploring the space of topic coherence measures," *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pp. 399–408, doi: 10.1145/2684822.2685324(2015).

[15] gensim.: "Topic coherence pipeline," https://radimrehurek.com/gensim/models/coherencemodel.html (2023).

[16] N. K. Tran, S. Zerr, K. Bischoff, C. Niederée, and R. Krestel.: "Topic cropping: Leveraging latent topics for the analysis of small corpora," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8092 LNCS, pp. 297–308, doi: 10.1007/978-3-642-40501-3_30(2013).

[17] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei.: "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, doi: 10.1162/tacl_a_00325(2020).

[18] A. A. Hidayat, R. Nirwantono, A. Budiarto, and B. Pardamean.: "BERT-based Topic Modeling Approach for Malaria Research Publication," pp. 326–331, doi: 10.1109/icimcis56303.2022.10017743(2023).

[19] "Similarity-based ranking of European tourism destinations leveraging Airbnb experiences and custom pre-trained TourBERT embeddings," (2022).

[20] N. Reimers and I. Gurevych.: "Sentence-BERT: Sentence embeddings using siamese BERT-networks," *EMNLP-IJCNLP Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 3982–3992, doi: 10.18653/v1/d19-1410(2019).

[21] L. McInnes, J. Healy, and J. Melville.: "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,", [Online]. Available: http://arxiv.org/abs/1802.03426(2018).

[22] R. J. G. B. Campello, D. Moulavi, and J. Sander.: "Density-Based Clustering Based on Hierarchical Density Estimates,", pp. 160–172(2018).

[23] T. Gelar and A. Nanda.: "Eksplorasi Pengembangan Korpus Pembicaraan Spontan pada Video

Instruksional Pertanian Perkotaan," *Journal of Software Engineering, Information and Communicaton Technology (SEICT)*, vol. 3, no. 1, pp. 111–120,doi: https://doi.org/10.17509/seict.v3i1.44548(2022).

[24] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf.: "FLAIR: An easy-to-use framework for state-of-the-art NLP," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Demonstrations Session*, pp. 54–59, (2019).