A Proposed Ensemble Approach for Searching Hacking News Semantically

DOI: https://doi.org/10.54654/isj.v2i22.1033

Do Ngoc Long, Nguyen The Hung*, Nguyen Trung Dung, Do Van Khanh, Nguyen Anh Tu, Pham Thi Bich Van

***Abstract*— Efficient search of hacking information has been a topic of great discussion in recent years. Many challenges are encountered when searching for this information. In particular, researchers may encounter unfamiliar and potentially challenging terms, ideas, tools, and other items that are unique to hacking. Effective comprehension of synonyms and polysemy is necessary. These reasons serve as the driving force behind our efforts to develop a productive method for semantic hacking information searches. Semantic search, using advanced NLP techniques, has transformed information retrieval by improving search result accuracy and relevance. Unlike traditional lexical methods, neural models like sentence-transformers handle synonyms and polysemy efficiently. However, processing time increases with model size. This paper proposes a novel ensemble semantic search (NESS) approach that aggregates mini or small neural embedding models, leveraging their distinct advantages. Evaluated on a dataset with over 300,000 Hacker News stories, our proposed method significantly enhances ranking quality and retrieval accuracy compared to existing techniques, while requiring half the processing time of the best-performing large model. The findings underscore the trade-offs between model complexity, retrieval accuracy, and processing efficiency, offering insights for optimizing semantic search systems.**

***Tóm tắt*— Tìm kiếm thông tin về các hoạt động tấn công một cách hiệu quả là chủ đề được thảo luận sôi nổi trong những năm gần đây. Nhiều thách thức gặp phải khi tìm kiếm những thông tin này. Đặc biệt, những khó khăn có thể gặp phải khi hiểu các thuật ngữ, ý tưởng, công cụ và một số mục không phổ biến chỉ dành riêng cho việc tấn công. Việc hiểu hiệu quả các từ đồng nghĩa và đa nghĩa là cần thiết. Những lý do này đóng vai trò là động lực thúc đẩy nỗ lực của nhóm tác giả nhằm phát triển một phương pháp hiệu quả để tìm kiếm thông tin hack ngữ nghĩa. Việc áp dụng các kỹ thuật xử lý ngôn ngữ tự nhiên hiện đại vào tìm kiếm theo ngữ nghĩa đã cải thiện đáng kể việc truy xuất thông tin bằng cách nâng cao độ chính xác và tính liên quan của kết quả tìm kiếm. So với các phương pháp truyền thống, các mô hình mạng nơron xử lý hiệu quả các từ đồng nghĩa và đa nghĩa. Tuy nhiên, mô hình càng lớn thì thời gian xử lý càng nhiều. Bài báo này đề xuất phương pháp tìm kiếm ngữ nghĩa (NESS) bằng cách kết hợp các mô hình nhúng nhỏ. Khi đánh giá trên tập dữ liệu chứa hơn 300.000 bản ghi từ trang Hacker News, NESS cải thiện đáng kể chất lượng xếp hạng và độ chính xác truy xuất so với các kỹ thuật hiện có, đồng thời thời gian xử lý giảm một nửa so với mô hình lớn có độ chính xác tốt nhất. Kết quả nhấn mạnh việc cân nhắc giữa độ phức tạp của mô hình, độ chính xác của kết quả và hiệu quả truy vấn, cung cấp những hiểu biết để tối ưu hóa các hệ thống tìm kiếm theo ngữ nghĩa.**

Keywords*— *semantic search; large language models; hacking news.

Từ khóa*— *tìm kiếm ngữ nghĩa; mô hình ngôn ngữ lớn; tin tức về hoạt động tấn công.

I. INTRODUCTION

As cyber threats evolve, ensuring robust cybersecurity measures is paramount to protecting sensitive data and maintaining user trust [1, 2]. Developing effective tools for information security search is essential to swiftly identify, analyze, and mitigate potential risks. These tools enable security professionals to stay ahead of emerging threats by providing timely and relevant information.

This manuscript was received on May 21, 2024. It was reviewed on June 17, 2024, revised on June 27, 2024 and accepted on September 23, 2024.

* Corresponding author

Cybersecurity researchers are interested in the numerous hacker communities that have emerged in recent years. Hackers gather in these networks to exchange resources and expertise related to cybercrime [3]. Hacking tools and stolen data are traded on underground markets in certain areas [4]. Nonetheless, there are other obstacles that researchers and practitioners must overcome in order to search hacker community contents. Comparing these communities to more conventional virtual communities, unconventional data is present in them. Researchers may not be aware of vocabulary, concepts, tools, or other hacker-specific topics that community members discuss [3, 4]. These motivations drive to develop a practical method for effectively looking for hacking information.

In terms of hacking information, various terminologies are used to describe the same concept [3]. For example, terms like InfoSec, network security, data breaches, and penetration testing, though related, often overlap in discussions. Semantic search can identify and match synonyms, ensuring comprehensive search results even if different terms are used. Sometimes users may not know the exact terminology for writing their query, which can cause keyword-based searches to fail. In addition, semantic search can effectively categorize content based on underlying topics rather than just keywords [5].

Semantic search represents a significant advancement in information retrieval, focusing on understanding the meaning behind user queries to deliver more relevant results. Unlike traditional keyword-based search systems, semantic search aims to comprehend the context, intent, and semantics of the query, leading to more accurate and contextually appropriate responses [6, 7]. This approach leverages advanced natural language processing (NLP) techniques [8, 9] and large language models (LLMs) [10, 11, 12], which have driven recent improvements in the field [13, 14, 15, 16].

The growing capabilities of LLMs, such as Bidirectional Encoder Representations from Transformers (BERT) [9], Generative Pre-Training (GPT) [10], and their successors, have significantly enhanced the ability to understand and process human language with remarkable precision. These models excel at capturing the nuanced relationships between words and concepts, making them ideal for powering semantic search engines. For instance, BERT's introduction marked a pivotal shift by enabling bidirectional context understanding, thereby improving the relevance and accuracy of search results. More recent models, like Large Language Model Meta AI (LLaMA) [12], continue to push the boundaries, offering more sophisticated semantic embeddings and contextual awareness.

In practical applications, semantic search has demonstrated its superiority across various domains, including healthcare, e-commerce, and legal research [17]. For example, generalist embedding models have shown to outperform domain-specific models in clinical semantic searches by effectively handling short-context queries and diverse expressions, thus challenging the traditional reliance on specialized models. This evolution underscores the potential of semantic search to transform how we interact with and retrieve information in both specialized and general contexts.

Current semantic search techniques are supported by large, high-accuracy embedding models, but they require powerful hardware and significant processing time. In this research, we evaluate the state-of-the-art semantic search methods and propose an efficient ensemble method for semantically scanning hacker news. The findings demonstrate that, in the realm of semantic search, our proposed method outperforms the assessed existing methods. The relevant field work is presented in the next section of the paper. Additionally, we present our suggested semantic search algorithm and introduce the experimental setup and outcomes. A conclusion and suggestions for further study are provided at the end.

II. Related work

The domain of semantic search has seen significant advancements in recent years, driven by the development of sophisticated models and

techniques to enhance the understanding and retrieval of relevant information. This section reviews some of the key contributions to the field, focusing on the evolution from traditional retrieval methods to advanced Transformer-based models like BERT and its variants.

1. Traditional retrieval methods

Traditional information retrieval systems have relied mainly on keyword-based approaches, such as TF-IDF and BM25 [18, 19]. TF-IDF is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents (corpus). The importance increases proportionally with the number of times a word appears in the document but is offset by the frequency of the word in the corpus. This helps in identifying words that are significant in a specific document but not common across the entire corpus. BM25 is an extension of the probabilistic information retrieval model and is considered one of the most effective ranking functions for document retrieval.

$$BM25(D,Q) = \sum_{i=1}^{n} IDF(q_i).\frac{f(q_i,D).(k_1+1)}{f(q_i,D)+k_1(1-b+b.\frac{|D|}{avgdl}} \quad (1)$$

Where, D - is the document; Q - is the query containing $q_1, q_2, ..., q_n$ terms; $f(q_i, D)$ is the frequency of the term q_i in document D; $|D|$ is the length of the document; $avgdl$ is the average document length in the collection; k_1, b are free parameters, usually chosen as $1.2 \leq k_1 \leq 2.0$ and $0 \leq b \leq 1$. $IDF(q_i)$ is the inverse document frequency weight of the term q_i. The $IDF(q_i)$ component is calculated as:

$$IDF(q_i) = \log(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1) \quad (2)$$

Where: N is the total number of documents in the collection; $n(q_i)$ is the number of documents containing the term q_i.

It calculates the relevance of a document based on the query terms it contains, adjusting for term frequency and document length. BM25 provides a more sophisticated approach than TF-IDF by incorporating term saturation and normalization.

Traditional retrieval methods, while effective to some extent, suffer from several limitations such as often missing relevant documents that use synonyms or related terms, lack of context and struggle with synonyms (different words with similar meanings) and polysemy (same word with different meanings).

2. Novel retrieval approaches based deep network models

With the advent of neural networks, information retrieval began to incorporate deep learning techniques [6, 8]. Word2Vec uses shallow, two-layer neural networks to produce word embeddings, which are dense vector representations of words in a continuous vector space. These embeddings capture semantic similarities between words based on their co-occurrence patterns in large corpora, enabling more effective retrieval by understanding the relationships between words. However, these models still have limitations in capturing context [6, 20].

3. Transformer models and BERT

The introduction of Transformer models by Vaswani et al. [11] marked a significant shift in NLP. Transformers use self-attention mechanisms to capture dependencies between words in a sequence, leading to more accurate contextual embeddings [11].

Building on this architecture, BERT was developed by Devlin et al. [9]. BERT's bidirectional approach enables it to understand context from both directions, significantly improving its ability to capture the meaning and relationships within text. BERT has been widely adopted in various NLP tasks, including semantic search, due to its robust performance in understanding query intent and document relevance.

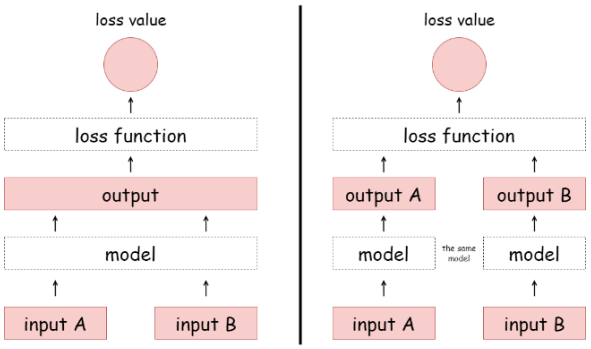

Figure 1. Cross-encoder architecture (left) and bi-encoder architecture (right)

In research [21], the *all-MiniLM-L6-v2* model is a compact and efficient transformer model designed for tasks like semantic search and text similarity. Fine-tuned on a large dataset using a self-supervised contrastive learning objective, this model is particularly useful for applications that require high performance with lower computational costs.

4. Advanced Variants and Extensions

Several extensions and variants of BERT have been proposed to further enhance its capabilities:

Sentence-BERT (SBERT) [22]: SBERT modifies the BERT architecture to generate semantically meaningful sentence embeddings, facilitating tasks such as semantic textual similarity and clustering. Given an anchor sentence a, a positive sentence p, and a negative sentence n, triplet loss tunes the network such that the distance between a and p is smaller than the distance between a and n:

$$\max\left(\| s_a - s_p \| - \| s_a - s_n \| + \in, 0\right) \quad (3)$$

with s_x the sentence embedding for $a / n / p$, $\| \cdot \|$ a distance metric and margin $\in$. Margin $\in$ ensures that s_p is at least $\in$ closer to s_a than s_n

Contextualized Late Interaction over BERT (ColBERT) [23]: ColBERT combines the efficiency of late interaction with the contextual embeddings from BERT, improving the performance of large-scale retrieval tasks by using multiple vectors for each text.

DistilBERT [24]: A smaller and faster version of BERT that retains most of its performance while reducing computational requirements. This model is particularly useful for resource-constrained environments.

5. BGE-M3 and Multi-Vector Retrieval

The BGE-M3 (Bert General Embeddings) model by the Beijing Academy of Artificial Intelligence (BAAI) [25] exemplifies the next step in the evolution of semantic search models. M3 distinguished for its versatility in Multi-Linguality, Multi-Functionality, and Multi-Granularity. It can support more than 100 working languages, leading to new state-of-the-art performances on multi-lingual and cross-lingual retrieval tasks. It also leverages a combination of dense, sparse (lexical), and multi-vector retrieval techniques to handle various retrieval scenarios effectively.

Multi-vector retrieval, a key feature of BGE-M3, enhances the representation of text by using multiple vectors to capture different semantic aspects. This approach allows for more granular matching between queries and documents, improving the accuracy and relevance of search results. ColBERT is another notable example that employs multi-vector retrieval to optimize the late interaction between query and document vectors.

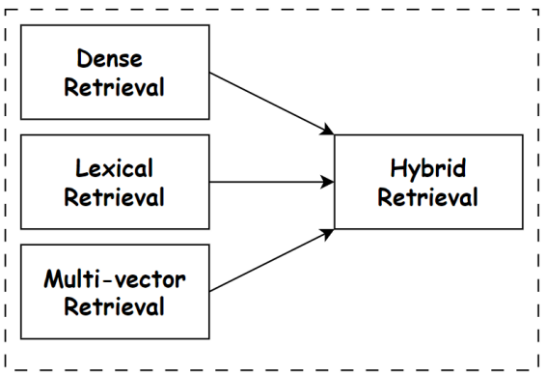

Figure 2. BGE-M3 multi-functionality

The BGE series includes three models for English: *bge-small-en-v1.5, bge-base-en-v1.5, and bge-large-en-v1.5*. These models are designed for English text processing tasks and offer a range of options balancing between computational efficiency and performance. These BGE models cater to various NLP

needs, allowing users to choose the model that best fits their specific application requirements in terms of balancing performance and resource efficiency.

In conclusion, the progression from traditional keyword-based methods to advanced Transformer-based models like BERT and its variants has significantly improved the performance and capabilities of semantic search systems. These advancements have enabled more accurate and context-aware retrieval, making semantic search a crucial component of modern information retrieval systems.

III. PROPOSED METHOD

In this study, we propose a novel ensemble learning approach in the field of semantic search to improve retrieval accuracy and ranking quality. Our method integrates multiple embedding models to leverage their unique strengths. The advantage of this approach is that it can reduce processing time while maintaining the same or higher level of accuracy as large models. This is achieved through the integration of small or aggregate models. We refer to the proposed approach as Novel Ensemble Semantic Search (NESS).

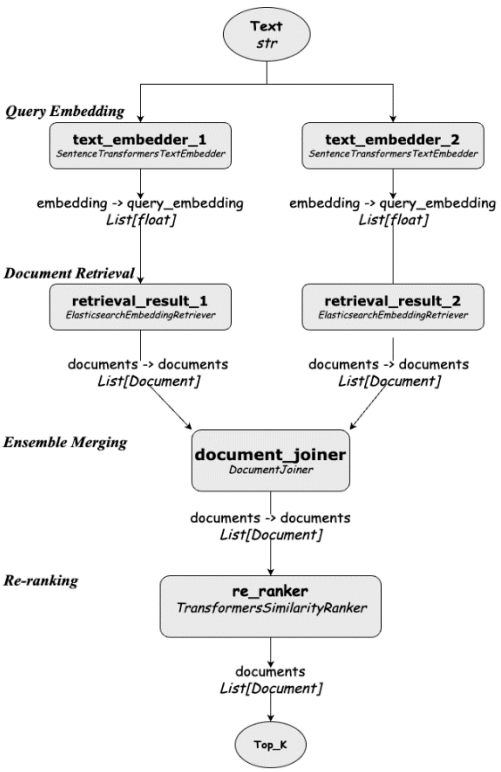

Figure 3. The architecture of NESS

The proposed pipeline consists of the following phases:

- *Phase 1: Query Embedding*

We use two pre-trained models, all-MiniLM-L6-v2 and BAAI/bge-small-en-v1.5. These models are used to embed the query separately. These models are chosen for their efficiency and effectiveness in generating dense vector representations.

- *Phase 2: Document Retrieval*

Each query embedding is used to retrieve a list of relevant documents from the corpus, resulting in two separate lists of documents. This results in two separate lists of documents, each ranked by the respective embedding model's similarity scores.

- *Phase 3: Ensemble Merging*

The retrieved document lists are combined into a single list. Documents appearing in both lists have their similarity scores aggregated, while unique documents retain their original scores from the respective model.

- *Phase 4: Re-ranking*

The combined document list is then re-ranked using *BAAI/bge-reranker-base*, a model specifically designed for fine-tuning the ranking of retrieved documents. This re-ranking process refines the initial retrieval by using a more sophisticated semantic understanding.

The proposed method aims to enhance the overall performance of semantic search by utilizing diverse model capabilities for initial retrieval and a specialized model for final ranking.

IV. EXPERIMENTAL DESIGN

This section details the experimental design employed to evaluate the performance of various sentence-transformer models in a semantic search pipeline. The design focuses on data preprocessing, model selection, indexing, query evaluation, and statistical analysis.

1. Data Collection and Preprocessing

Data Source: The dataset used in this study is the training set of kerinin/hackernews-stories, containing 313317 stories from Hacker News

[26]. The dataset is a comprehensive collection of stories and comments extracted from Hacker News, a popular social news website focusing on computer science and entrepreneurship.

Preprocessing: The dataset was preprocessed by the team to remove erroneous records due to errors during the scraping process. This was done to ensure data quality. Filtering of erroneous records involved identifying and removing entries that were incomplete or contained corrupted data. The cleaned dataset containing 278072 records was then prepared for indexing by converting each story into a format suitable for Elasticsearch import. Each document is processed by splitting the "Text" field into segments of 500 words with an overlap of 32 words to maintain context continuity between segments.

Figure 4 and Figure 5 are information about the length (in words) of the two data columns *Title* (Question) and *Text* (Answer) of the dataset. These figures are compiled based on our statistical analysis of the dataset. These statistical findings allow us to reasonably design parameters like *split_size* and *chunk_size*.

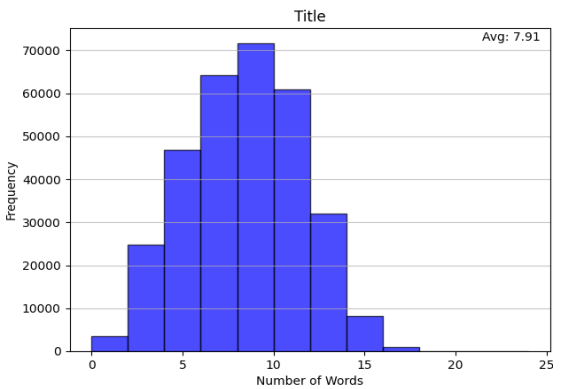

Figure 4. Word Number of Title or question field

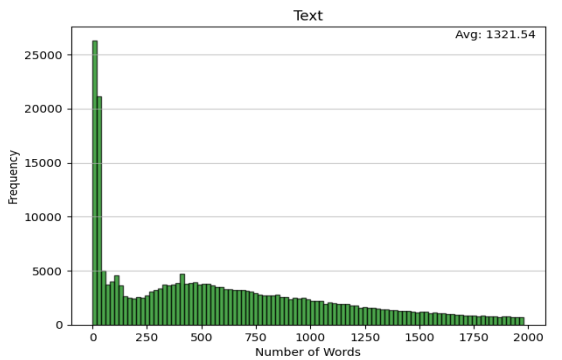

Figure 5. Word Number of Text or answer field

2. Evaluated selected models

In this paper, we evaluate our NESS with the *all-MiniLM-L6-v2 [21], bge-small-en-v1.5, bge-base-en-v1.5, and bge-large-en-v1.5 [25]* model. Table 1 shows the setups and their parameters.

TABLE 1. PARAMETERS OF SENTENCE-TRANSFORMER MODELS

Model	Embedding Dimensions	Model size (M)
all-MiniLM-L6-v2	384	23
bge-small-en-v1.5	384	33
bge-base-en-v1.5	768	109
bge-large-en-v1.5	1024	335
NESS (our method)	384	23 & 33

Each model processes the same dataset, generating embeddings for the segmented text. These embeddings are then ingested into their corresponding Elasticsearch instance. This process ensures that the embeddings are indexed and stored in a manner that facilitates efficient retrieval and analysis.

3. Indexing into Elasticsearch

Elasticsearch setup: An Elasticsearch instance was set up to serve as the backend for indexing and retrieving documents. The cleaned and preprocessed dataset was indexed into Elasticsearch using the Haystack framework.

Indexing process and parameters: Each document (story) was encoded using the selected sentence-transformer models to generate dense vector representations. These vector embeddings were then indexed into Elasticsearch to facilitate efficient similarity searches.

4. Evaluation method

Query Selection: We randomly selected 1,000 questions from the dataset to serve as queries for evaluating the search pipeline. This random sampling ensures a diverse set of queries, which helps in assessing the robustness and generalizability of the search models.

Evaluation Metrics

Time Measurements: We recorded the time taken to process 1,000 queries to assess the

efficiency of the search pipeline. We ran the experiment 30 times with 1,000 queries each and calculated the mean and standard deviation of the processing times.

Top-N Accuracy: The relevance of search results was evaluated using Top-1, Top-5, and Top-10 accuracy metrics. This involves checking if the correct answer or a highly relevant document appears within the top N results for each query. We conducted the experiments 30 times with 1,000 queries, reporting the mean and standard deviation for these accuracy metrics as percentages (%).

5. Implementation

Four identical Elasticsearch instances are set up to serve as the backend for indexing and retrieving documents, each running on different ports. Each instance is dedicated to one of the four models for storing their respective embeddings, ensuring that each model's performance can be independently evaluated without interference from the others. The instance configurations are as follows: Instance 1 on port 9200, Instance 2 on port 9201, Instance 3 on port 9202, and Instance 4 on port 9203.

The experimental setups are executed on a supercomputer with an Intel Core i7 CPU, 32GB of RAM, an RTX 4090 24GB GPU, and an Ubuntu 20.04 operating system.

V. Result and discussion

In this work, we examined the suggested methodology, NESS, using four pre-existing neural models on two metrics: accuracy (%) and processing time. The evaluation's results are displayed in Table 2. The values in the table are presented as mean ± standard deviation.

TABLE 2. THE COMPARISON RESULTS AMONG MODELS THROUGH RUNNING 30 TIMES WITH 1,000 RANDOM QUERIES

Model	Processing time (ms)	Top 1 (%)	Top 5 (%)	Top 10 (%)
all-MiniLM-L6-v2	12±1	37±1	57±2	63±2
bge-small-en-v1.5	15±4	42±2	60±2	65±2
bge-base-en-v1.5	56±10	43±2	62±2	67±2
bge-large-en-v1.5	404±34	45±2	65±2	71±1
NESS (Proposed method)	**181±16**	**48±2**	**68±2**	**73±1**

The experimental results demonstrate that our ensemble method significantly improves the retrieval performance compared to using individual models alone in term of the accuracy metric. When compared to merely the highest level of the models, which is roughly 45%, 65 %, and 71% for the large model, NESS achieves an accuracy of 48%, 68% and 73%, respectively. In addition, it also requires less than half of the processing time of the model with the best accuracy. In conclusion, by aggregating small embedding models effectively, our proposed method captures a broader range of semantic nuances, resulting in more accurate and relevant search outcomes while significantly reducing processing time.

Furthermore, the results illustrate the trade-offs between retrieval accuracy, processing efficiency, and model complexity, offering guidance for improving semantic search systems. It is evident that larger models take longer to generate search results even when they attain higher accuracy.

VI. Conclusion

In this paper, we first provide a brief overview of semantic search strategies and then examine current neural methods. We also proposed a novel ensemble semantic search (NESS), evaluated on a Hacker News stories dataset. The results showed that our proposed method, which integrates multiple mini or small embedding models to leverage their strengths, significantly improves ranking quality and retrieval accuracy compared to existing techniques. The outcomes highlight the balance between retrieval precision, computational efficiency, and model complexity, guiding the enhancement of semantic search systems.

Future research will explore integrating additional models and applying this method to other domains to further validate its effectiveness.

REFERENCES

[1] Sun, Nan, et al (2023). Cyber threat intelligence mining for proactive cybersecurity defense: a survey and new perspectives. IEEE Communications Surveys & Tutorials 2023.

[2] Thakur, Manikant. Cyber security threats and countermeasures in digital age. Journal of Applied Science and Education (JASE) 4.1 (2024): 1-20.

[3] Benjamin, Victor, and Hsinchun Chen. "Developing understanding of hacker language through the use of lexical semantics." 2015 IEEE International Conference on Intelligence and Security Informatics (ISI). IEEE, 2015.

[4] Li, Ying, et al. "NEDetector: Automatically extracting cybersecurity neologisms from hacker forums." Journal of Information Security and Applications 58 (2021): 102784.

[5] Satyapanich, Taneeya, Tim Finin, and Francis Ferraro. "Extracting rich semantic information about cybersecurity events." 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. *Efficient estimation of word representations in vector space*. CoRR, abs/1301.3781.

[7] Mitra, B., & Craswell, N. (2018). *An introduction to neural information retrieval*. Foundations and Trends in Information Retrieval, 13(1), 1-126.

[8] Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., & Poria, S. (2023). *A review of deep learning techniques for speech processing*. Information Fusion, 101869.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv:1810.04805.

[10] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. *Improving language understanding by generative pre-training*. Technical report, OpenAI.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. *Attention is all you need*. In Advances in Neural Information Processing Systems (pp. 5998-6008).

[12] Hugo Touvron, Thibaut Lavril, Xavier Martinet, et al. 2023. *LLaMA: Open and Efficient Foundation Language Models*. arXiv:2302.13971.

[13] Tom Kenter, Alexey Borisov, and Maarten de Rijke. 2016. Siamese CBOW: *Optimizing word embeddings for sentence representations*. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 941-951).

[14] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: *A conversational question answering challenge*. Transactions of the Association for Computational Linguistics, 7, 249-266.

[15] Brandon Nye, JinJin Li, Ramya Patel, et al. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (pp. 197-203).

[16] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. *A deep relevance matching model for ad-hoc retrieval*. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2251-2259).

[17] JD Prater. *AI-Powered Semantic Search: Everything You Need to Know*. https://www.graft.com/blog/the-future-is-semantic-transforming-search-in-the-age-of-ai (2023).

[18] Trotman, Andrew, Antti Puurula, and Blake Burgess (2014). *Improvements to BM25 and language models examined*. Proceedings of the 2014 Australasian Document Computing Symposium (ADCS 2014) (pp. 58-65).

[19] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: *BM25 and Beyond*. Foundations and Trends in Information Retrieval 3, 4 (2009), 333-389.

[20] Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Dou, Z., & Wen, J. 2023. *Large Language Models for Information Retrieval: A Survey*. arXiv, abs/2308.07107.

[21] Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. *MiniLMv2: Multi-Head Self-Attention Relation Distillation for Compressing Pretrained Transformers*. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2140–2151, Online. Association for Computational Linguistics.

[22] Reimers, Nils and Iryna Gurevych. 2019. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural

Language Processing (pp. 3982-3992).

[23] Khattab, Omar, and Matei Zaharia. *Colbert: Efficient and effective passage search via contextualized late interaction over bert*. Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. 2020 (pp. 39–48).

[24] Sanh, Victor et al. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv abs/1910.01108.

[25] Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. 2024. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv, abs/2402.03216.

[26] Ryan Michael. *kerinin/hackernews-stories*. huggingface.co/datasets/kerinin/hackernews-stories.

ABOUT THE AUTHORS

Do Ngoc Long

Workplace: Institute 486, Command 86, Vietnam.

Email: longdn2k@gmai.com

Education: He graduated from the Military Technical Academy in Information Technology in 2023.

Recent research direction: cyber security, natural language processing, image processing.

Tên tác giả: **Đỗ Ngọc Long**

Cơ quan công tác: Viện Nghiên cứu 486, Bộ Tư lệnh 86, Việt Nam.

Email: longdn2k@gmai.com

Quá trình đào tạo: Tốt nghiệp kỹ sư ngành Công nghệ thông tin tại Đại học Kỹ thuật Lê Quý Đôn năm 2023.

Hướng nghiên cứu hiện nay: An ninh mạng, xử lý ngôn ngữ tự nhiên, xử lý ảnh.

Nguyen The Hung

Workplace: Institute 486, Command 86, Vietnam.

Email:hungnguyenthe1211@gmai.com

Education: He received a PhD and and MSc in Computer Science at the University of New South Wales, Canberra, Australia in 2023 and 2018, respectively. He received his B.E degree in Information Technology from Military Technical Academy, Vietnam in 2010.

Recent research direction: cyber security, trusted artificial intelligence, generative artificial intelligence, natural language processing, image processing.

Tên tác giả: **Nguyễn Thế Hùng**

Cơ quan công tác: Viện Nghiên cứu 486, Bộ Tư lệnh 86, Việt Nam

Email: hungnguyenthe1211@gmai.com

Quá trình đào tạo: Nhận bằng thạc sĩ và tiến sĩ chuyên ngành Khoa học máy tính tại Đại học New South Wales, Úc vào các năm 2018 và 2023. Tốt nghiệp kỹ sư ngành Công nghệ thông tin tại Đại học Kỹ thuật Lê Quý Đôn năm 2010.

Hướng nghiên cứu hiện nay: an ninh mạng, trí tuệ nhân tạo đáng tin cậy, trí tuệ nhân tạo tạo sinh, xử lý ngôn ngữ tự nhiên, xử lý ảnh.

Nguyen Trung Dung

Workplace: Institute 486, Command 86, Vietnam.

Email: ntdtoanud2011@gmail.com

Education: He received a PhD in the field of mathematical basis for information technology in 2019. He received a Master and B.E degree in Information Technology at Military Technical Academy, Vietnam in 2002 and 2008, respectively.

Recent research direction: cyber security, natural language processing, image processing.

Tên tác giả: **Nguyễn Trung Dũng**

Cơ quan công tác: Viện Nghiên cứu 486, Bộ Tư lệnh 86, Việt Nam.

Email: ntdtoanud2011@gmail.com

Quá trình đào tạo: Nhận bằng tiến sĩ chuyên ngành Cơ sở toán học cho Công nghệ Thông tin tại Viện KHCN Quân sự năm 2019. Tốt nghiệp thạc sĩ và kỹ sư ngành Công nghệ thông tin tại Đại học Kỹ thuật Lê Quý Đôn vào các năm 2002 và 2008.

Hướng nghiên cứu hiện nay: An ninh mạng, xử lý ngôn ngữ tự nhiên, xử lý ảnh.

Do Van Khanh

Workplace: Institute 486, Command 86, Vietnam.

Email: khanhdovanit@gmai.com

Education: He received a MSc in Computer Science at Warsaw University of Technology, Warsaw, Poland in 2021. He received his B.E degree in Information Technology from Military Technical Academy, Vietnam in 2013.

Recent research direction: dense-based information retrieval systems, domain-specific QA systems, and speech recognition systems.

Tên tác giả: **Đỗ Văn Khánh**

Cơ quan công tác: Viện Nghiên cứu 486, Bộ Tư lệnh 86, Việt Nam

Email: khanhdovanit@gmai.com

Quá trình đào tạo: Nhận bằng thạc sĩ chuyên ngành Khoa học máy tính tại Đại học Công nghệ, Warsaw, Ba Lan năm 2021. Tốt nghiệp kỹ sư ngành Công nghệ

thông tin tại Đại học Kỹ thuật Lê Quý Đôn vào các năm 2013.

Hướng nghiên cứu hiện nay: hệ thống truy xuất thông tin dựa trên mật độ dày đặc, hệ thống QA, hệ thống nhận dạng giọng nói.

Nguyen Anh Tu

Workplace: Institute 486, Command 86, Vietnam.

Email: nguyenanhtu789@gmail.com

Education: He received a PhD in Informatics and Computing, MSc and B.E in Automation and Control at Tomsk Polytechnic University, Tomsk, Russia in 2019, 2015 and 2013 respectively.

Recent research direction: large language model, knowledge acquisition/extraction and cybersecurity.

Tên tác giả: **Nguyễn Anh Tú**

Cơ quan công tác: Viện Nghiên cứu 486, Bộ Tư lệnh 86, Việt Nam

Email: nguyenanhtu789@gmail.com

Quá trình đào tạo: Nhận bằng tiến sĩ trong Tin học và máy tính, thạc sĩ và đại học chuyên ngành Tự động hóa và điều khiển tại Đại học Bách khoa Tomsk, Nga vào các năm 2019, 2015, và 2013.

Hướng nghiên cứu hiện nay: mô hình ngôn ngữ lớn, thu thập/khai thác kiến thức và an ninh mạng.

Pham Thi Bich Van

Workplace: Institute of Information and Comunication Technology, Le Quy Don Technical University, Vietnam.

Email: vanptb@lqdtu.edu.vn

Education: Van completed her Bachelor of Information and Technology and Master of Information Systems at the LeQuyDon Technical University in 2010 and 2014 resprectively. She received her PhD in Software Engineering at the University of New South Wales, Canberra, Australia in 2020.

Recent research direction: software metrics, software quality, cyber security.

Tên tác giả: **Phạm Thị Bích Vân**

Cơ quan công tác: Viện Công nghệ thông tin và truyền thông, Đại học Kỹ thuật Lê Quý Đôn, Việt Nam

Email: vanptb@lqdtu.edu.vn

Quá trình đào tạo: Nhận bằng Tiến sĩ chuyên ngành Kỹ sư phần mềm tại Đại học New South Wales, Úc năm 2020. Tốt nghiệp thạc sĩ và kỹ sư ngành Công nghệ thông tin tại Đại học Kỹ thuật Lê Quý Đôn vào các năm 2014 và 2010.

Hướng nghiên cứu hiện nay: độ đo phần mềm, chất lượng phần mềm, an ninh mạng.