Using Ontology to Analyze Sentiment of Comments on Vietnamese Social Media

Nguyen Viet Hung[1,*,†], Nguyen Anh Quan[1,†], Nguyen Van Vu[1,†], Phan Thi Yen[2,†], Nguyen Hai Binh[1,†] and Nguyen Thi Thuy Nga[1]

[1]*Faculty of Information Technology, East Asia University of Technology, Bacninh, Vietnam*
[2]*Scientific Management Department, East Asia University of Technology, Bacninh, Vietnam*

Abstract

Recently, there has been a growing trend in studies that employ ontology-based methods to analyze sentiment in social media comments in Vietnam. Ontology, a model comprising concepts, attributes, and relationships, serves as a knowledge reference framework for expressing emotions in comments. This approach enhances understanding of how Vietnamese individuals convey emotions on platforms such as YouTube, Facebook, and others. In contrast to traditional sentiment analysis methods, ontology aims to achieve more detailed and accurate sentiment analysis by leveraging semantic connections between concepts. Therefore, this paper proposes: (1) employing ontology for sentiment analysis in Vietnamese social media, (2) collecting and preprocessing comment data from popular platforms in Vietnam, (3) utilizing ontology to assign sentiment labels (positive, negative) to comments, (4) analyzing sentiment patterns and trends in comments, and (5) evaluating the performance of ontology-based methods versus traditional sentiment analysis. The findings of this study contribute to advancing social data analysis techniques and offer insights into user behaviors on Vietnamese social media platforms. Experiments also show that the proposed method achieves the best performance compared to other methods, with an accuracy of up to 0.8657 and an F1 score of up to 0.9174.

Keywords

Analyze Sentiment, Social Media, Ontology, Vietnamese, Opinion

1. Introduction

Technological advancements in recent years have profoundly changed people's lives in the physical world. With technology becoming more advanced and accessible, fundamental developments such as analysis, evaluation, and commentary have been integrated. Given that comments encompass a wide array of issues, including sentiment, emotion, and interaction, analyzing comments has become essential in today's digital age. Sentiment analysis (SA) is intricately tied to the rapid technological advancements in the real world. It stands out as one of the most dynamic research domains within Natural Language Processing (NLP), owing to its significant potential applications in both business and society [1, 2, 3]. This calls for the development of new evaluation models and presents considerable challenges. While these models have seen widespread use in recent years, there remains room for improvement in this trend. Unfortunately, these models are trained on various architectures, pre-trained data, and preprocessing steps, leading to inconsistencies and errors in systems. All studies compare and evaluate performance using both the monolingual PhoBERT and the ViT5 model [1]. Furthermore, the study [1] stated that it is the first to investigate the performance of fine-tuning Transformer-based models on five datasets with different domains and scales for Vietnam's SA task. Considering

Joint Workshop on Knowledge Diversity and Cognitive Aspects of KR (KoDis/CAKR) co-located with the 21st International Conference on Principles of Knowledge Representation and Reasoning (KR2024), November 2–8, 2024 Hanoi, Vietnam
*Corresponding Author: hungnv@eaut.edu.vn
†These authors contributed equally.
hungnv@eaut.edu.vn (N. V. Hung); 20220844@eaut.edu.vn (N. A. Quan); 20220839@eaut.edu.vn (N. V. Vu); yenpt@eaut.edu.vn (P. T. Yen); binhnh@eaut.edu.vn (N. H. Binh); ngantt@eaut.edu.vn (N. T. T. Nga)
https://nvhung278.blogspot.com/ (N. V. Hung); https://eaut.edu.vn/ (N. A. Quan); https://eaut.edu.vn/ (N. V. Vu); https://eaut.edu.vn/ (P. T. Yen); https://eaut.edu.vn/ (N. H. Binh); https://eaut.edu.vn/ (N. T. T. Nga)
0000-0002-9767-6749 (N. V. Hung); 0009-0007-1058-0925 (N. A. Quan); 0009-0007-9185-080X (N. V. Vu); 0009-0006-8974-8462 (P. T. Yen); 0009-0009-6581-5386 (N. H. Binh); 0009-0009-3579-051X (N. T. T. Nga)

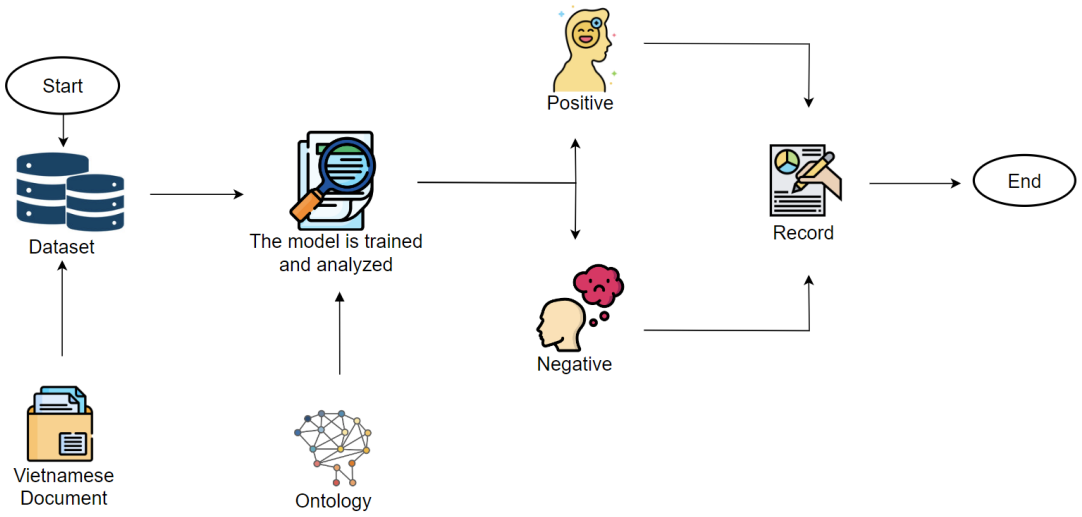

Figure 1: Architecture System Overview

cultural factors is crucial during application, as misinformation can directly impact the training of the model [4, 5, 6, 7, 8].

To effectively analyze Vietnamese comments on social media platforms [4, 8], we propose a comprehensive system that integrates natural language processing and machine learning techniques. This system aims to classify the sentiment of Vietnamese comments as either positive or negative, offering valuable insights into user opinions and emotions, as depicted in Figure 1.

Therefore, to analyze the sentiment of comments on Vietnamese social networks, we propose the following listed implementation procedures:

- Using Ontology for sentiment analysis in Vietnamese social media.
- Collecting and preprocessing comment data from popular Vietnamese social media platforms.
- Using Ontology to assign sentiment labels (positive, negative, neutral) to comments.
- Analyzing sentiment patterns and trends in comments.
- Comparing the performance of the Ontology-based method to traditional sentiment analysis.

This study's findings contribute to expanding social data analysis techniques and provide practical insights into user behavior on Vietnamese social media platforms. These insights can be directly applied to enhance social media strategies and user engagement.

The rest of this paper is organized as follows: Section 2 discusses related work, Section 3 describes the proposed model, and Section 4 evaluates performance. Finally, Section 5 discusses our findings and pending issues.

2. Related Work

This research relates to our investigation on chatbots, social media comment analysis, and ontology. The study focuses on the impact of managerial relationships, social media, and business attitudes on the financial performance of small and medium-sized enterprises (SMEs) in Vietnam [9]. The report outlines some theoretical advances and provides useful recommendations for practitioners in Vietnamese SMEs who want to increase productivity and efficiency.

Motivated by GLUE's success, they provide the Social Media Text Classification Evaluation (SMTCE) benchmark, which consists of a set of models and datasets covering a wide range of SMTC tasks. They apply and evaluate a range of BERT-based [10] multilingual (mBERT, XLM-R, and DistilmBERT) and monolingual (PhoBERT, viBERT, vELECTRA, and viBERT4news) models for tasks in the SMTCE benchmark using the suggested benchmark. Monolingual models produce state-of-the-art results on all

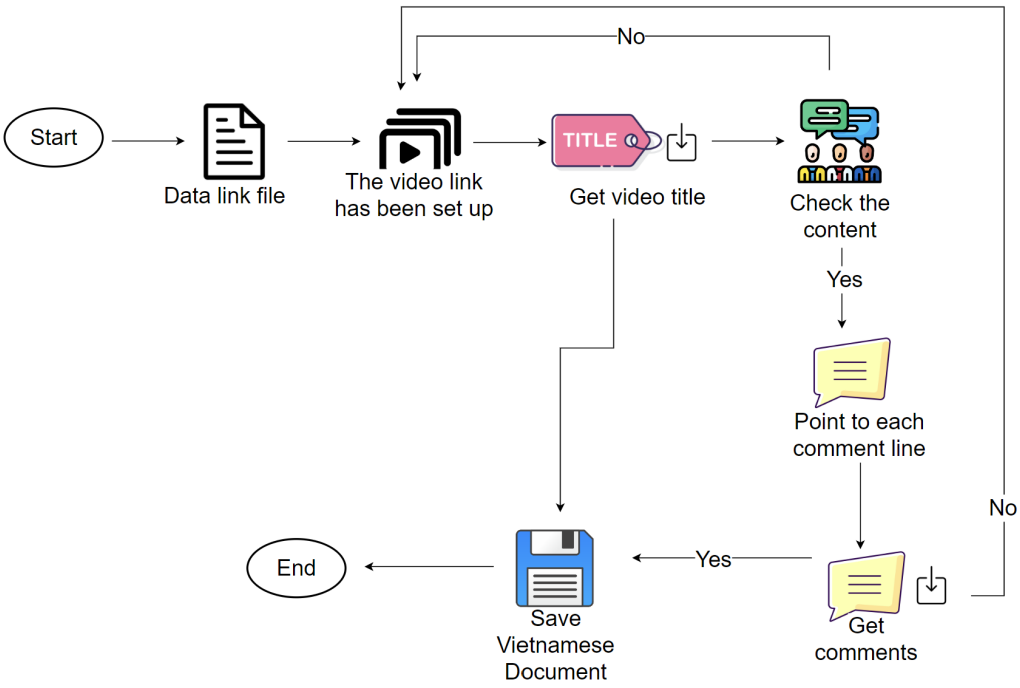

Figure 2: Automatic comment collection and analysis of our proposed System

text classification tasks and outperform multilingual models. It offers a neutral evaluation of BERT-based models that are both monolingual and multilingual using the standards, which would be helpful for future research on BERTology in the Vietnamese language.

The study of [11] examines customer perceptions of Vietnamese hotel services in general and aspects of hotel services by combining natural language rules and inferential statistics; they analyze to understand customer experiences with the hotel industry recovering from the pandemic and thereby provide what customers want to enhance the customer experience.

The purpose of [12] is to investigate the factors leading to social commerce adoption in Vietnam. The participants of this study were 447 social networking website users in Vietnam. The results identify important antecedents that influence Vietnamese consumers' propensity to participate in social commerce. These findings have implications for research and practical applications in understanding social commerce adoption in emerging economies.

One of the negative side effects of more people using social networking sites is an increase in rude and nasty language aimed at other members. Because of this, reviewing tagged comments that have been filtered by categorization systems may become challenging for human moderators. In an effort to solve this problem, [13] presents the ViHOS (Vietnamese Hate and Offensive Spans) dataset. They also offer comprehensive annotation rules and definitions of hateful and offensive spans in Vietnamese comments.

Around the world, a lot of individuals utilize social media for education and amusement. Furthermore, as several foreign language specialists and academics across the world have demonstrated, using this kind of technology helps students learn foreign languages, including English, Chinese, French, Japanese, and so forth. The purpose of the study [14] is to see how students feel about expanding their vocabulary in English through social media use.

Globally, a large number of individuals utilize social media for learning and amusement. Additionally, the usage of this kind of technology aids with children's foreign language acquisition, as noted by several academics and specialists in foreign languages from across the globe. The study's [15] goal was to find out how students felt about utilizing social media to expand their vocabulary in English. Van Lang University (VLU), in Vietnam, used a blend of quantitative and qualitative methodologies. Fifteen

of the 154 students who participated in semi-structured interviews had surveys completed on them by different majors. Their research showed a connection between negative rejection-related stresses and negative FOMO ratings, as well as a relationship between FOMO scores and worse overall quality of life and increased depression symptoms.

They present a technique in this article for gauging social tension in various Russian regions by examining user posts on the social network Vkontakte (VK) [16, 17]. They created a tool to gather postings from VK members that expressed unfavorable opinions regarding prevalent societal problems like inflation, corruption, and unemployment. Using this tool, they were able to compile data on the quantity of these postings made over specific timeframes and examine user profiles in general.

3. Proposed Model

In this study, we aim to explore Vietnamese comments on social media. We identify and analyze elements of these platforms to develop an effective self-learning support system. The process involves two main stages: data collection and processing, followed by the analysis and evaluation of the comments in Figure 2.

First, we set up the working environment and necessary tools. Next, Vietnamese document datasets, including articles, comments, and other documents related to students' learning on social networks, will be transferred to our database system. This process requires meticulous preparation to collect the data thoroughly and accurately using programming techniques.

Data collection and processing: we gather data from various sources, including popular social media platforms. The collected data is then cleaned and normalized to meet system requirements. This processing involves removing noise, reformatting the data, and organizing it into appropriate categories. We use ontologies to train the model after transferring the data to the system. Ontologies help identify concepts and the relationships between these concepts in the learning and education domain. Incorporating ontologies during model training allows the system to understand and analyze the data more deeply and accurately.

We begin the data analysis once the data is prepared and the system model is trained. Our system employs natural language processing (NLP) techniques and machine learning algorithms to analyze the sentiment of the comments in the dataset. This includes semantic analysis, keyword identification, and sentiment ratings for each comment.

After the system completes the analysis process, the findings will be displayed as sentiment classifications and used for further steps, including understanding the context of the comments. Finally, the program will store the comments and analysis results in the system. This ensures that all data is securely stored and provides a foundation for future updates and analysis.

The sentiment analysis module is at the core of the proposed system [18], which integrates natural language processing techniques and machine learning algorithms. It plays a crucial role in enabling the system to understand and analyze linguistic data, thus providing users with accurate and valuable assessments.

3.1. Data collection

Many studies have demonstrated that self-study on social networking platforms is highly practical. We collected comments from YouTube to assess users' attitudes and emotions regarding self-study using social networks and analyzed them using sentiment analysis (SA). Our automated comment data collection system is illustrated in detailed in Algorithm 1 as follows:

- **System Initialization**: This stage involves starting and preparing the system to execute the subsequent tasks.
- **File Transmission**: The system receives a specific file containing a list of video URLs.
- **Link Retrieval**: The system will retrieve each video URL from the provided file in a predetermined order to ensure a sequential process.

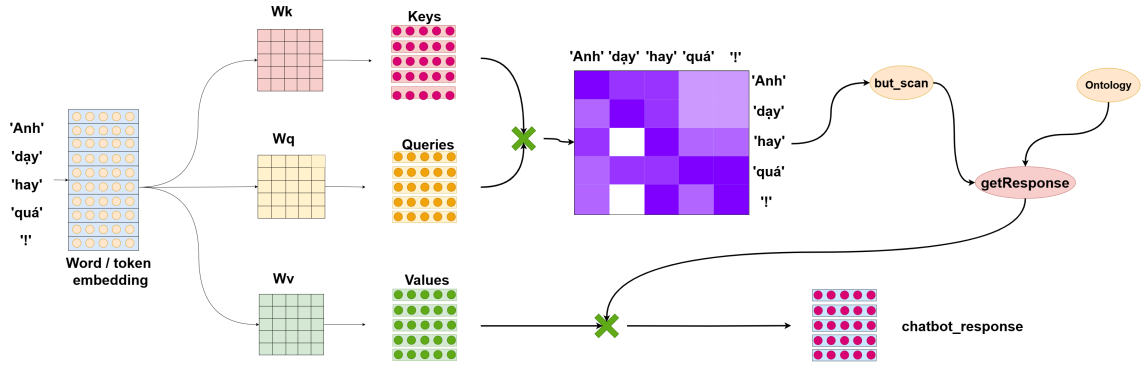

Figure 3: Start the data analysis system

Algorithm 1 Data collection

Input: $List\_Video$
Parameter: L, L_i, i, n
Output: $comment$, $title$

```
L = read_file(List_Video)
for i from 0 to n - 1 do
    L_i = read_line(L,i)
    access_link(L_i)
    title = get_title(L_i)
    comments = get_comments(L_i)
    if title and comments then
        for comment in comments do
            write_to_VietNamDocument(title, comment)
        end for
    else
        continue
    end if
end for
```

- **Video Title Extraction**: The system will extract the title of each video from the provided links.
- **Video Content Verification**: The system will assess the content of each video to determine its reliability.
- **Comment Section Identification**: If the video contains content, the system will locate and retrieve the section containing video comments.
- **Comment Extraction**: The system will extract all comments from the designated comment section.
- **Comment Count Storage**: In the final and crucial step, upon successful extraction of comments, the system will accurately record the number of comments in the Vietnam Document dataset.
- **Program Termination**: The program will conclude after completing the aforementioned steps.

In summary, this process involves collecting and analyzing information from Vietnamese-related videos and their comments using data retrieval, natural language processing, and data storage techniques.

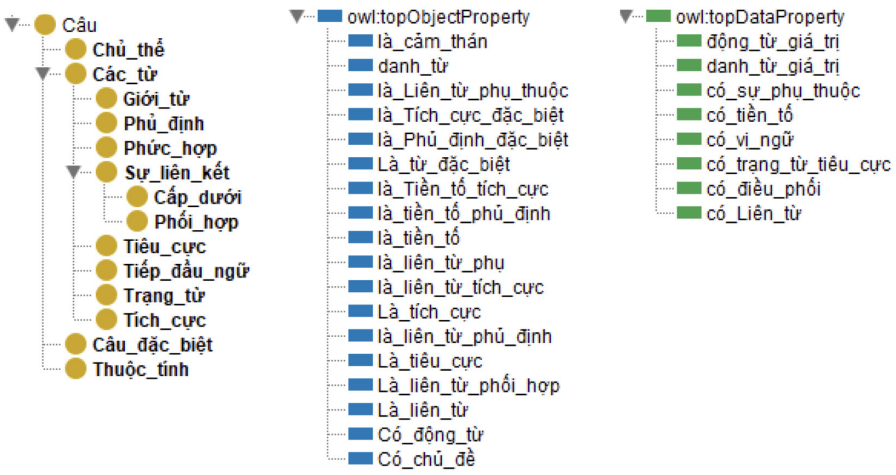

Figure 4: Flowchart depicting our ontology

3.2. Data processing system

After the data undergoes raw processing and the removal of any irregular characters, it is input into the word analysis system. Below is an outline of how our proposed system operates:

- **Step 1**: When words are input into the system, it first identifies the sentence structure and then segments the sentence in Figure 3. These segments are subsequently examined and filtered to eliminate special characters. This filtering step is crucial in text processing to ensure that data is cleansed before analysis, retaining only alphabetic, numeric, and whitespace characters in the input string. After cleaning the string, the process proceeds by breaking down the sanitized sentence into individual words, a process known as tokenization. Tokenization converts a text string into a list of words, with each word representing a distinct unit of meaning in the text. This step prepares the data for more advanced semantic analysis and text-processing tasks.
- **Step 2**: The system evaluates the sentiment of sentences by analyzing their context to determine whether they convey a positive or negative meaning. It employs word scanning techniques to identify and filter out sentences with negative sentiment. This process enhances the accuracy of comments and predictions. Next, the system extracts key words from sentences and analyzes them by cross-referencing with its ontology. The ontology is structured as a binary tree, which helps classify text into two primary branches: positive and negative. Each branch represents a different type of intent. The left branch corresponds to negative features, while the right branch pertains to positive elements. Finaly, this ontology provides a detailed and structured description of terms and their relationships. It aids the system in understanding the semantics of the text more thoroughly and applying logical inference rules effectively.
- **Step 3**: Once the ontology is constructed and trained, the system employs a binary search algorithm to query words within the user's sentence against the ontology. This search method efficiently retrieves terms from binary tree data structures. Based on the results of this ontology search, the system can determine whether the user's sentence carries a positive, negative, or neutral sentiment. Corpus comparison and processing algorithms are utilized to conduct this assessment, evaluating the sentence's structure and content in relation to key terms defined in the ontology in Figure 4.

4. Performance Evaluation

In this paper, we assess the effectiveness of the proposed method by analyzing 3.739 Vietnamese comments. Additionally, we compare our method with three existing methods that analyze English

Table 1
Performance of the proposed method compared to other approaches

Values	ITEAI	ACCLE	BCSAO	The proposed
Accuracy	0.7622	0.5027	0.7853	0.8657
F1-score	0.7381	0.0082	0.7602	0.9174

comments. These reference methods are a chatbot system designed to analyze English opinions (referred to as BCSAO), a chatbot for changing lifestyles in education (referred to as ACCLE), and an interactive transport inquiry AI chatbot (referred to as ITEAI). Below is a summary of these reference methods:

- **ITEAI** [19]: Similar to the ACCLE approach, this method develops a chatbot system that queries users about their current location and final destination. The design analyzes the user's query and fetches relevant data from the database. It provides comprehensive information, ensuring individuals can safely reach their desired destination.
- **ACCLE** [20]: The author suggests a Chatbot system to enhance teacher and student collaboration. In this system, students submit text-based questions to the Chatbot, which uses natural language processing and deep learning technologies to process the data and respond to the students. However, this system is limited to use within schools and does not analyze the respondents' emotions.
- **BCSAO** [18]: This method resembles the one we propose. However, while their data processing mainly relies on programming techniques, our approach goes further in applying them. Creating an ontology categorizes sentences and performs a more detailed analysis for each specific topic, followed by automated separation using individual models and optimizing comment sentences.

In conclusion, to assess the proposed method's performance against the evaluation methods, we utilize the following formulas according to [21] to calculate Accuracy and F1-score.

Our study aims to assess the accuracy of refer models. Formula 1 determines the ratio of correct predictions to the total number of predictions made. The formula for calculating accuracy is presented below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Additionally, we recognize that accuracy is only sometimes the best metric for all scenarios. It can be particularly limited in cases of data imbalance, where one class significantly outweighs the other. Therefore, we also compute other metrics, such as precision (formula 2), recall (formula 3), and F1-score (formula 4). The formulas are as follows:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$\textit{F1-score} = 2 * \frac{Pre * Rec}{Pre + Rec} \tag{4}$$

Where:

- TP: The model predicts 1, and the actual value is also 1.
- TN: The model predicts 0, and the actual value is also 0.
- FN: The model predicts 0, but the actual value is 1.
- FP: The model predicts 1, but the actual value is 0.

Our system outperforms the three methods listed in Table 1. The proposed method consistently achieves a higher percentage than the other methods. The ACCLE method shows a low average result of 0.5027, while the ITEAI and BCSAO methods achieve 0.7622 and 0.7853, respectively. The proposed method reaches the lowest level at 0.8657.

5. Conclusions

In this study, we developed a social media comment evaluation model using ontology techniques to classify Vietnamese comments as positive or negative. We collected Vietnamese comments from video platforms using Python programming, enhancing our ability to analyze user sentiment. This approach is particularly beneficial for businesses in Vietnam that deal with a high volume of customer comments. Our method significantly outperforms existing ones, achieving an accuracy improvement of up to 0.0804 compared to others.

On the other hand, we also evaluate educational teaching videos by assessing their quality and effectiveness in conveying content through video lectures. Although comments have been collected from these videos, future research will aim to gather a broader and more diverse range of comments to enhance the study's scope.

For future work, it is essential to explore how the proposed framework can be applied across different languages, cultures, and age groups. Additionally, understanding how the system identifies and interprets irony or sarcasm in comments will be a key area of focus.

6. Acknowledgments

This research is funded by the East Asia University of Technology (EAUT).

References

[1] D. Van Thin, D. N. Hao, N. L.-T. Nguyen, Vietnamese sentiment analysis: An overview and comparative study of fine-tuning pretrained language models, ACM Transactions on Asian and Low-Resource Language Information Processing 22 (2023) 1–27.

[2] B. Liu, Sentiment analysis and opinion mining, Springer Nature, 2022.

[3] R. Jindal, K. Seeja, S. Jain, Construction of domain ontology utilizing formal concept analysis and social media analytics, International Journal of Cognitive Computing in Engineering 1 (2020) 62–69.

[4] V. T. Dũ, et al., Social media use by vietnamese journalists: Current status and solutions, Revista de Gestão Social e Ambiental 18 (2024) e06270–e06270.

[5] F. Sufi, A sustainable way forward: Systematic review of transformer technology in social-media-based disaster analytics, Sustainability 16 (2024) 2742.

[6] N. V. Hung, T. Q. Loi, N. T. Huong, T. T. T. Hang, T. T. Huong, Aafndl-an accurate fake information recognition model using deep learning for the vietnamese language, Информатика и автоматизация 22 (2023) 795–825.

[7] W. Graterol, J. Diaz-Amado, Y. Cardinale, I. Dongo, E. Lopes-Silva, C. Santos-Libarino, Emotion detection for social robots based on nlp transformers and an emotion ontology, Sensors 21 (2021) 1322.

[8] T. Le, V.-H. Nguyen, T. Ho, A model of discovering customer insights in tourism sector approach to vietnamese reviews analytics, in: 2022 9th NAFOSTED Conference on Information and Computer Science (NICS), IEEE, 2022, pp. 205–210.

[9] A. Nguyen, P. v. Nguyen, H. Do, The effects of entrepreneurial orientation, social media, managerial ties on firm performance: Evidence from vietnamese smes, International Journal of Data and Network Science 6 (2022) 243–252.

[10] L. T. Nguyen, K. Van Nguyen, N. L.-T. Nguyen, Smtce: A social media text classification evaluation benchmark and bertology models for vietnamese, arXiv preprint arXiv:2209.10482 (2022).
[11] H. T. T. Nguyen, T. X. Nguyen, Understanding customer experience with vietnamese hotels by analyzing online reviews, Humanities and Social Sciences Communications 10 (2023) 1–13.
[12] R. Cutshall, C. Changchit, H. Pham, D. Pham, Determinants of social commerce adoption: An empirical study of vietnamese consumers, Journal of Internet Commerce 21 (2022) 133–159.
[13] P. G. Hoang, C. D. Luu, K. Q. Tran, K. Van Nguyen, N. L.-T. Nguyen, Vihos: Hate speech spans detection for vietnamese, arXiv preprint arXiv:2301.10186 (2023).
[14] T. Pham Manh, V. Nguyen, T. Cao Thi Xuan, Vietnamese students' perceptions of utilizing social media to enhance english vocabulary: A case study at van lang university, Pham, MT, Nguyen, TTV, & Cao, TXT (2023). Vietnamese Students' Perceptions of Utilizing Social Media to Enhance English Vocabulary: A Case Study at Van Lang University. International Journal of TESOL & Education 3 (2023) 79–111.
[15] V. A. T. Dam, N. G. Dao, D. C. Nguyen, T. M. T. Vu, L. Boyer, P. Auquier, G. Fond, R. C. Ho, C. S. Ho, M. W. Zhang, Quality of life and mental health of adolescents: Relationships with social media addiction, fear of missing out, and stress associated with neglect and negative reactions by online peers, Plos one 18 (2023) e0286766.
[16] D. Donchenko, N. Ovchar, N. Sadovnikova, D. Parygin, O. Shabalina, D. Ather, Analysis of comments of users of social networks to assess the level of social tension, Procedia Computer Science 119 (2017) 359–367.
[17] I. Kozitsin, A. Chkhartishvili, A. Marchenko, D. Norkin, S. Osipov, I. Uteshev, V. Goiko, R. Palkin, M. Myagkov, Modeling political preferences of russian users exemplified by the social network vkontakte, Mathematical Models and Computer Simulations 12 (2020) 185–194.
[18] H. V. Nguyen, N. Tan, N. H. Quan, T. T. Huong, N. H. Phat, Building a chatbot system to analyze opinions of english comments, Информатика и автоматизация 22 (2023) 289–315.
[19] M. Dharani, J. Jyostna, E. Sucharitha, R. Likitha, S. Manne, Interactive transport enquiry with ai chatbot, in: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 1271–1276. doi:10.1109/ICICCS48265.2020.9120905.
[20] E. Kasthuri, S. Balaji, A chatbot for changing lifestyle in education, in: 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 1317–1322. doi:10.1109/ICICV50876.2021.9388633.
[21] H. Nguyen, T. N. Dao, N. S. Pham, T. L. Dang, T. D. Nguyen, T. H. Truong, An accurate viewport estimation method for 360 video streaming using deep learning, EAI Endorsed Transactions on Industrial Networks and Intelligent Systems 9 (2022) e2. URL: https://publications.eai.eu/index.php/inis/article/view/2218. doi:10.4108/eetinis.v9i4.2218.