An Ontology-based Sentiment Analysis Approach to Discovering Hidden Affected Objects

Le Anh Phuong[1], Nguyen Minh Duc[2], Nguyen Dinh Hoa Cuong[3], Nguyen Thi Huong Giang[1]
[1] Department of Computer Science, University of Education, Hue University, Hue
[2] Department of Business Information System, University of Economics, Hue University, Hue
[3] School of Engineering and Technology, Hue University, Hue
Tác giả liên hệ: Nguyen Thi Huong Giang, Email: nthgiang@hueuni.edu.vn
Ngày nhận bài: 29/08/2022, ngày sửa chữa: 17/11/2022, ngày duyệt đăng: 25/11/2022
Định danh DOI: 10.32913/mic-ict-research-vn.v2022.n2.1146

Tóm tắt: Nowadays, the topic of sentiment analysis has attracted a wide range of artificial intelligence technologies such as Natural Language Processing (NLP), neural network. Long Short-Term Memory (LSTM), a deep neural network, has proven the effiency in the task of sentiment classification. However, there is a shortage of semantic knowledge from such classification results. This paper introduces a sentiment analysis approach based on a knowledge model to discover hidden affected objects, which combines ONtology and SentiMent technology, named ONSEM. In order to confirm the efficiency of the ONSEM framework, an experiment on a sentiment analysis dataset was developed and then yielded promising experimental results.

Từ khóa: *Sentiment analysis, ontology, long short-term memory, semantic reasoning, hidden affected object.*

Title: **Phương pháp phân tích quan điểm dựa trên Ontology để khám phá các đối tượng ẩn**

Abstract: Ngày nay, nhiều công nghệ được sử dụng để nghiên cứu chủ đề phân tích quan điểm như trí tuệ nhân tạo, xử lý ngôn ngữ tự nhiên (NLP), mạng nơ ron thần kinh. Bộ nhớ ngắn hạn dài (LSTM), một mạng lưới thần kinh sâu, đã chứng minh sự hiệu quả trong việc phân loại quan điểm. Tuy nhiên, việc dựa vào kết quả phân loại này bị thiếu hụt về kiến thức ngữ nghĩa. Bài báo này giới thiệu một phương pháp phân tích quan điểm dựa trên mô hình tri thức để khám phá những đối tượng ẩn có liên quan, được đặt tên là ONSEM. Để xác nhận sự hiệu quả của ONSEM, một thí nghiệm trên bộ dữ liệu phân tích quan điểm đã được phát triển và đã gặt hái những kết quả đầy triển vọng.

Keywords: *Phân tích quan điểm, Ontology, trí nhớ ngắn hạn dài, suy luận ngữ nghĩa, đối tượng ảnh hưởng bị ẩn.*

I. INTRODUCTION

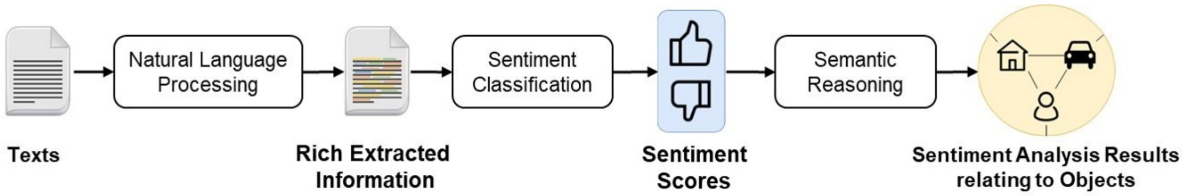
Sentiment analysis employing NLP techniques has been widely applied in many fields, such as market research, finance, politics [1]. In the recent decade, deep learning has reached a big achivement on sentiment classification [2]. Especially, LSTM, an architecture of recurrent neural network (RNN), has shown an outstanding performance because of the ability to handle length texts.

Although LSTM has been adopted in various studies of sentiment classification, the methods of handling knowledge in accordance with sentiment scores were not studied thoroughly. Specifically, related hidden objects affected by sentiment evaluations were not discovered. Ontology, the core of the Semantic Web, has been well-known for the ability of representing knowledge models and semantic reasoning [3]. This study introduces an ontology-based sentiment analysis approach in order to discover objects hiddenly affected by users' reviews, named ONSEM. An experiment on testing the performance of the proposed framework was deployed. The experimental results showed the adequacy of the ONSEM framework.

The rest of this paper is organized as follows: Section II presents relevant studies. Section III introduces the ONSEM framework in detail while the experiment is introduced in Section IV. Lastly, Section V gives the conclusions and suggests future research directions.

II. RELATED WORK

The full elaboration of sentiment analysis studies is out of the scope of this paper. Readers are suggested to find the comprehensively related information in the following survey [4].



Hình 1. Conceptual block diagram of the ONSEM framework.

Sentiment analysis has strong connection to NLP due to the task of preprocessing text data. NLP transforms text to input vectors for classification models. The typical NLP tasks include: to kenization, POS tagging, lemmatization, stemming, stop word removal, and named entity recognition. Typical approaches of sentiment analysis employing NLP can be found in recent studies [5–9].

Within the domain of classification, in recent years, the deep learning approaches have been widely applied to the problem of sentiment analysis because of their high performances of prediction [10]. The literature has been introduced many different deep learning approaches including: (i) weighted Convolutional Neural Networks (CNN) for sentiment analysis of texts [5]; (ii) multiple CNNs composing of CNN, DCNN, MVCNN, and VDCNN for classifying sentiments of Bangla text [6]; (iii) hierarchy CNN for analyzing multi-entity sentiment analysis [7]; (iv) Bi-directional RNN for overcoming the limitations of semantic-information loss and irrelevant shared features [8]; (v) LSTM model for sentiment analysis of hotel reviews [9].

Beside the aformentioned classification models, rule-based classifiers and ontology have also been applied to sentiment analysis. For example, a Semantic Web Rule Language (SWRL) rule-based classifier and ontology was applied to analyzing sentiment of fake news [11]. Sentiment analysis was considered under the view of concept levels with the support of ontology [12]. Sentiment analysis could be used to predict customer interest by combining with e-commerce ontology and fuzzy method [13].

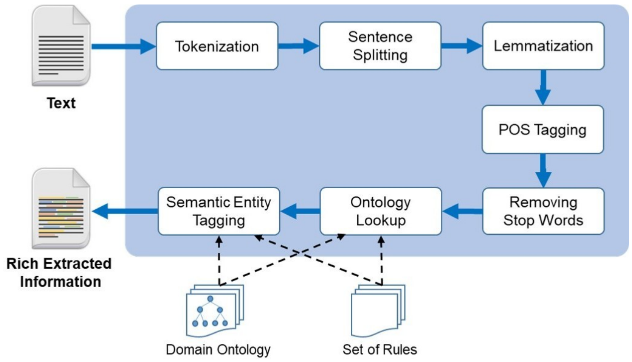
III. METHODOLOGY

1. The ONSEM framework

Figure 1 shows an overview of the ONSEM framework, which consists of three modules including *natural language processing*, *sentiment classification* and *semantic reasoning*. Each of these modules is described in detail as follows.

2. Natural language processing

The first module aims to capture the syntactic and semantic information from the input texts. As shown in Figure 2,



Hình 2. NLP pipeline.

the NLP module includes 7 steps where some steps are supported by NLP tool such as Stanford CoreNLP [14].

Firstly, *tokenization* splits texts into a sequence of tokens, each storing types of text such as word, number, punctuation mark or symbol, etc. Next, *sentence splitting* determines sentences in the texts utilizing the sentence boundary indicators such as periods, question marks, exclamation points. *Lematization* is the task of deriving the root form of each word in inflectional form or in derivational form. The major step is *POS tagging* where each token is labelled with its part-of-speech such as noun, verb, adverb. Insignificant words (e.g. 'a', 'an') are then filtered out by the step of *Removing stop words*. The two last steps utilizes a domain ontology and a set of rules to extract semantic entities from the input texts. While the major information extracted are processed at the next module of *sentiment classification*, the annotation results of semantic entities are further employed at the module of *semantic reasoning*.

3. Sentiment classification

Before being applied sentiment classification, tokens are vectorized by the TF-IDF method [15] (Term Frequency-Inverse Document Frequency) to measure how relevant they are to the input text. The TF-IDF score for the word t in the document d from the set of document D is presented in (1), (2) and (3). Specifically, $freq(t, d)$ is the frequency of the word t in the document d and $count(d \in D : t \in d)$

shows the number of document d in the set of document D containing the word t.

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D) \quad (1)$$

$$tf(t, d) = \log(1 + freq(t, d)) \quad (2)$$

$$idf(t, D) = \log\left(\frac{N}{count(d \in D : t \in d)}\right) \quad (3)$$

In this study, LSTM network [16] was selected for the classification model. Figure 3 shows the LSTM structure at time step t. Specifically, i, o, f denote the input gate, the output gate and the forget gate, respectively. C_t is the memory cell at time t. The inputs of input gate i_t, output gate o_t and forget gate f_t are vector x_t that the network receives at time t and its previous hidden state h_{t-1}.

The LSTM component is formalized as follows:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f), \quad (4)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \quad (5)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o), \quad (6)$$

$$\hat{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c), \quad (7)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \hat{c}_t, \quad (8)$$

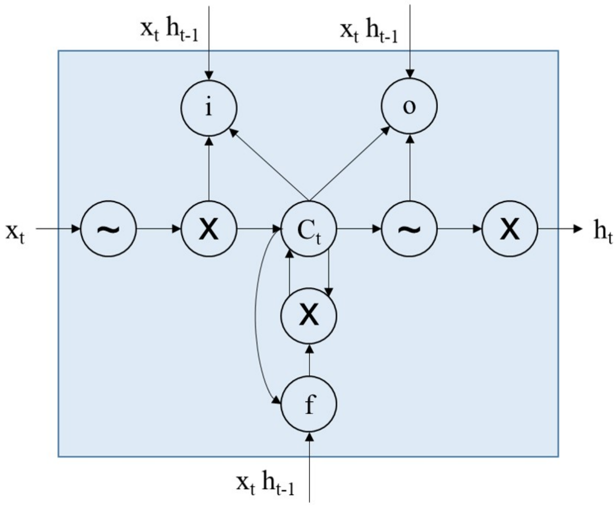
$$h_t = o_t \circ \tanh(c_t), \quad (9)$$

where σ, $\circ$, respectively denote a logistic sigmoid function and element-wise multiplication; f_t, i_t, o_t, c_t, h_t are the forget gate, the input gate, the output gate, the memory cell and hidden state at time-step t. W, U, b are respectively two weights matrices and a bias vector. Within this scope, while the forget gate decides to forget unnecessary information, the input gate directs what new information should be stored in the memory cell. At the final stage, the output gate determines the amount of information in the memory cell that should be exposed. The LSTM model could remember significant information over time steps by this mechanism.

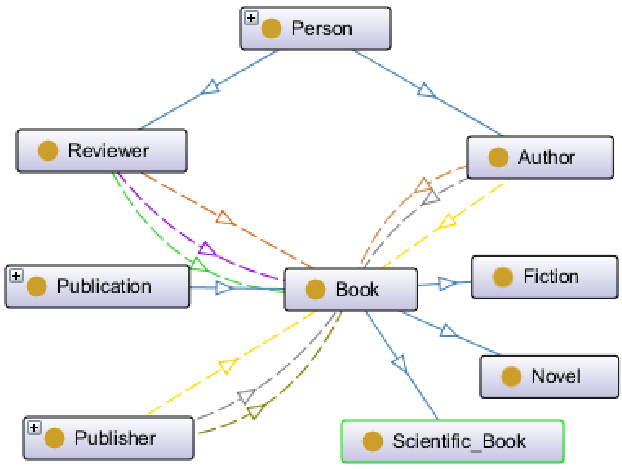
4. Semantic reasoning

The ONSEM semantic reasoning module composes of a domain ontology and its SWRL base. Theoretically, an ontology of domain D is defined as $O_D = \langle C_D, R_D, P_D, I_D \rangle$ where C_D, R_D, P_D, and I_D are the sets of concepts, relations, data properties, and instances, respectively. Within the scope of this paper, a domain ontology of book reviews was developed for the purpose of demonstrating the efficiency of ONSEM framework. An excerpt of this ontology is illustrated in Figure 4.

Based on this ontology, a SWRL rule base was developed in order to discover hidden objects affected by users' reviews. Specifically, results of sentiment analysis found by LSTM model are populated into the ONSEM ontology. Then, the semantic reasoning engine implements its SWRL rules to discover hidden affected objects. Some examples SWRL rules are presented in Table I.



Hình 3. LSTM structure.



Hình 4. An excerpt of the ONSEM ontology.

Bảng I
EXAMPLES OF SWRL RULES

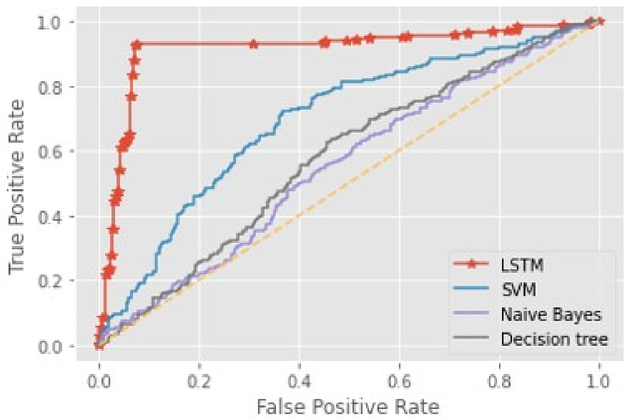
No.	Rule
1	Book(?b) ∧ Author(?a) ∧ Review(?r) ∧ has_Author(?b, ?a) ∧ has_Negative_Review(?b,? r) → negative_Affect_By(?a, ?r)
2	Book(?b) ∧ Author(?a) ∧ Review(?r) ∧ has_Author(?b, ?a) ∧ has_Positive_Review(?b,? r) → positive_Affect_By(?a, ?r)

IV. EXPERIMENT

In order to demonstrate the efficiency of ONSEM framework, a sentiment analysis dataset was retrieved from the Book Genome dataset [17]. Specifically, a tabular dataset containing book title, authors, review text, and rating was

Bảng II
AUC SCORES.

Model	AUC scores
LSTM	**0.914**
SVM	0.670
Naïve Bayes	0.553
Decision tree	0.580

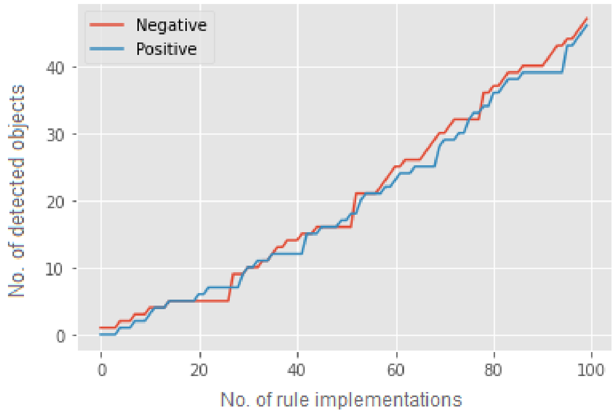


Hình 5. ROC curves of classifiers.

retrieved from JSON-format files of Book Geneom dataset. In which, the original 5-star scale rating was mapped to the binary scale including: (i) negative (1-3 star rating); and (ii) positive (4-5 star rating). Then, any record, which has empty review text, is ignored. Based on this standardized dataset, a 2000-record dataset, which has balance rating labels, was selected and used in our experiments. Additionally, the objects of this dataset (e.g. authors, books) were also populated into the ONSEM ontology as instances. The training set and test set were formed up following the rule of 70%-30%. All of classification models were trained with 10-fold cross-validation method.

The experiments aimed at: (i) confirming the performance of the LSTM model of ONSEM framework; and (ii) demonstrating the reasoning ability of ONSEM's semantic module in discovering objects affected by users' reviews. For the former, the three well-known machine learning models including SVM, Naïve Bayes and decision tree were used as the baseline methods. The performances of these classification models were visualized by ROC curves in Figure 5 and summarized by their AUC scores in Table II.

As showed in Figure 5, the ROC curve of LSTM-based model dominated those of the other models. This domination was also supported by the LSTM AUC score, which is greater than AUC scores of the other classifiers.



Hình 6. Discovering affected objects by reasoning.

The experimental results figured out that LSTM-model outperformed three baseline models.

In order to measure the efficiency of semantic reasoning module, we implemented the process of sentiment analysis and semantic reasoning over the test set and counted the discovered objects affected by users' reviews. The cumulative results of negatively affected objects and positively affected objects are depicted in Figure 6. This promising result confirmed the ability of discovering hidden information by using reasoning process of the ONSEM approach.

V. CONCLUSION

In this study, an ontology-based sentiment analysis approach discovering hidden affected objects was introduced. The ONSEM framework including three modules of natural language processing, sentiment classification and semantic reasoning was described in detail. An experiment on a sentiment analysis dataset was deployed to demonstrate the feasibility of ONSEM framework. For further works, a hybrid approach of LSTM and another deep learning technique will be considered to gain a better result.

TÀI LIỆU THAM KHẢO

[1] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams engineering journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[2] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM*, vol. 56, no. 4, pp. 82–89, 2013.

[3] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 34–43, 2001.

[4] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—a review of research topics, venues, and top cited papers," *Computer Science Review*, vol. 27, pp. 16–32, 2018.

[5] A. Ghorbanali, M. K. Sohrabi, and F. Yaghmaee, "Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks," *Information Processing & Management*, vol. 59, no. 3, p. 102929, 2022.

[6] N. R. Bhowmik, M. Arifuzzaman, and M. R. H. Mondal, "Sentiment analysis on bangla text using extended lexicon dictionary and deep learning algorithms," *Array*, vol. 13, p. 100123, 2022.

[7] C. Gan, L. Wang, and Z. Zhang, "Multi-entity sentiment analysis using self-attention based hierarchical dilated convolutional neural network," *Future Generation Computer Systems*, vol. 112, pp. 116–125, 2020.

[8] Y. Cai, Q. Huang, Z. Lin, J. Xu, Z. Chen, and Q. Li, "Recurrent neural network with pooling operation and attention mechanism for sentiment analysis: A multi-task learning approach," *Knowledge-Based Systems*, vol. 203, p. 105856, 2020.

[9] P. F. Muhammad, R. Kusumaningrum, and A. Wibowo, "Sentiment analysis using word2vec and long short-term memory (lstm) for indonesian hotel reviews," *Procedia Computer Science*, vol. 179, pp. 728–735, 2021.

[10] J. Joseph, S. Vineetha, and N. Sobhana, "A survey on deep learning based sentiment analysis," *Materials Today: Proceedings*, 2022.

[11] D. Vidanagama, A. Silva, and A. Karunananda, "Ontology based sentiment analysis for fake review detection," *Expert Systems with Applications*, vol. 206, p. 117869, 2022.

[12] A. Khattak, M. Z. Asghar, Z. Ishaq, W. H. Bangyal, and I. A. Hameed, "Enhanced concept-level sentiment analysis system with expanded ontological relations for efficient classification of user reviews," *Egyptian Informatics Journal*, vol. 22, no. 4, pp. 455–471, 2021.

[13] R. Karthik and S. Ganapathy, "A fuzzy recommendation system for predicting the customers interests using sentiment analysis and ontology in e-commerce," *Applied Soft Computing*, vol. 108, p. 107396, 2021.

[14] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60.

[15] A. Aizawa, "An information-theoretic perspective of tf–idf measures," *Information Processing & Management*, vol. 39, no. 1, pp. 45–65, 2003.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] D. Kotkov, A. Medlar, A. Maslov, U. R. Satyal, M. Neovius, and D. Glowacka, "The tag genome dataset for books," in *ACM SIGIR Conference on Human Information Interaction and Retrieval*, 2022, pp. 353–357.

AUTHOR'S SUMMARY

Le Anh Phuong is teaching and researching at the Department of Computer Science, Hue University of Education. The author graduated from the Department of Mathematics, Hue University of Education in 1996; received Master and PhD degree in 2001, 2013 at Ha Noi University of Science and Technology. He has been qualified as Associate Professor since 2017. His research interests include Computational Logic, Knowledge Representation and Uncertainty Argument, Hedge Algebra, Descriptive Logic and Semantic Web, Decision Support Systems.

Email: laphuong@hueuni.edu.vn

Nguyen Minh Duc is teaching and researching at the Department of Business Information System, University of Economics, Hue University. The author graduated in 2014 at Department of Economic Information Systems, Hue University; gained Master and PhD degree in 2017, 2021 at Inje University, Korea. His current research focuses are on Data Mining, Text Engineering, Knowledge Management, Ontology, Natural Language Processing.

Email: nguyenminhduc@hueuni.edu.vn

Nguyen Dinh Hoa Cuong is teaching and researching at the School of Engineering and Technology, Hue University. The author graduated in 2002, received Master degree in Informatics in 2006 from University of Science, Hue University and received PhD in Computer Science in 2016 at Khon Kaen University. His research interests are about Data Mining, Text Mining, Knowledge Management, Recomender systems.

Email: ndhcuong@hueuni.edu.vn

Nguyen Thi Huong Giang is teaching and researching at the Department of Computer Science, Hue University of Education. The author graduated in 2002 and received Master Informatics in 2006 from University of Science, Hue University. Her research interests are about Data Mining, Knowledge Management, Ontology.

Email: nthgiang@hueuni.edu.vn