Capturing Contextual Factors in Sentiment Classification: An Ensemble Approach

THIEN KHAI TRAN\textsuperscript{1,2,3} AND TUOI THI PHAN\textsuperscript{1,2}
\textsuperscript{1}Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam
\textsuperscript{2}Vietnam National University at Ho Chi Minh City, Ho Chi Minh City, Vietnam
\textsuperscript{3}Ho Chi Minh City University of Foreign Languages and Information Technology (HUFLIT), Ho Chi Minh City, Vietnam

Corresponding author: Thien Khai Tran (8141216@hcmut.edu.vn)

This research was funded by Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, under Grant number BK-SDH-2020-8141216.

ABSTRACT Sentiment classification is a crucial task in sentiment analysis, and has received significant attention from researchers. Previous studies have focused on using several techniques to solve this problem. However, to the best of our knowledge, none of these works has fully investigated the exploitation and the manipulation of contextual information in the text, or taken advantage of the combined power of state-of-the-art models. In this paper, we propose an effective ensemble learning model for the sentiment classification problem. In our system, the contextual information in the text is fully captured by integrating rule-based methods and other state-of-the-art deep learning models. We found that the combination of word embedding representation and the attention mechanism, along with pre-defined rules and specific-domain sentiment dictionaries are helpful in dealing with numerous valence-shifting cases. Although the computational cost of the proposed system is higher than those of certain other algorithms, this system obtains better results than other approaches when tested on three different datasets.

INDEX TERMS Sentiment classification, contextual valence shifters, ensemble learning, deep learning, attention mechanism.

I. INTRODUCTION

Sentiment analysis has been one of the most exciting areas of natural language processing since 2002 [1]. The primary task of sentiment analysis is to automatically determine the value or semantic orientation of a document [2], which refers to a measure of subjective opinion that indicates a polarity (positive, negative, or neutral) and the strength of words, phrases, sentences, or documents. Sentiment analysis has been studied through two approaches [1]: (i) as a machine learning method that implements the construction of text classifiers by selecting features and algorithms that match the labeled data, and (ii) as a semantic-oriented approach that includes the calculation of the overall polarization through the semantic orientation of words or phrases in the text. Numerous tools and applications were developed to exploit user-generated content on the web. However, the performance of these systems is not high due to the complexity of natural languages. Many studies have shown that sentiment analysis is a more complex problem than merely subjective text classification [3]. These approaches remain ineffective in dealing with some linguistic phenomena, especially with a contextual valence shifter.

Of the two solutions to the sentiment analysis problem mentioned above, the machine learning methodology is dominant because of increasingly abundant training data as well as good quality data sources that belong to a specified application domain within the problem’s goals. For example, to classify sentiments from hotel reviews, training data is typically sourced from tourism-hotel reviews, which can lead to the highest level of accuracy. However, machine learning methods often have difficulty in scenarios that cause value shifting due to the context. In contrast, based on the rules for calculating the sentiment values of words, phrases, sentences, and paragraphs, semantic orientation approaches can solve many cases with contextual valence shifters. Unfortunately, defining the rules that cover all situations of linguistic data is not feasible. This challenge is especially true for the Vietnamese language as data sources for sentiment analysis are not yet fully available, for example, the ontology WordNet [4] or a dictionary, such as SentiWordNet [5] for Vietnamese.
In this paper, we propose a novel sentiment classification methodology that adopts ensembles of algorithms to learn on various sub-datasets while focusing on contextual features. We mine the contextual features in text as follows. (1) We select a word embedding to model the text and the relationship between the context and the target word. Text representation is the key to obtain better performance for classifiers, including sentiment classifiers. To refactor the words into machine learning algorithms, text data must be converted to vector representations. Although one-hot representation is a simple method to implement, this approach leads to a loss of contextual information of the text and results in a high-dimensional space. (2) With attention mechanisms showing promising results in natural language processing tasks [6]–[8], including sentiment analysis [9], [10], our model applies this technique to capture all words in the context. (3) Domain-specific dictionaries are used to assign contextual scores to each sentiment word in the dataset, which are necessary for the context-based learners in our ensemble model.

The main contributions of our paper are as follows:

- We propose a novel ensemble model for sentiment classification by integrating many sub-models learned on various sub-datasets.
- Contextual factors in linguistics are fully exploited. A rule-based method captures contextual valence-shifting, and deep learning models improve the adaptive ability of the feature extraction tasks. We also adopt word embedding representation and attention mechanisms along with domain-specific sentiment dictionaries.
- Experiments on multiple datasets using our proposed model suggests improved results compared to the state-of-the-art models.

The remainder of the paper is organized as follows. Section II presents related preliminaries and previous research. Section III reviews the proposal of our model, and Section IV describes our experiments and evaluations. Finally, Section V presents our conclusion and introduces directions for future research.

II. PRELIMINARIES AND RELATED WORK

A. CONTEXTUAL VALENCE SHIFTERS

In addition to describing events objectively, texts often convey information about the attitudes of writers, or participants of a described event. This emotional attitude is expressed through the choice and the arrangement of words. Although some words always show positive or negative polarity, others are likely to have contextual valence changes due to the influence of nearby words as well as the organization of the words in the text. Cases of valence-shifting include the following:

- Negation: The most common type of valence-shifting, as in “She doesn’t like this laptop”, uses the negative “doesn’t” to overturn the emotional value of the word “like”. We can list many negation words in Vietnamese like không, not, chẳng, chẳng bao giờ, chẳng bao giờ không, chẳng bao giờ không, không, cắm, bắn, bắn, bắn, bắn.
- Defective verbs: Words used to express ability, certainty, permission, request, competence, proposal, order, obligation, or advice represent defective verbs that are always paired with another verb that has semantic content and modifiers for the verb. In cases where a sentence appears with a defective word, the sentimental value of the sentence is often negated with the value of the sentiment word or phrase accompanied by the verb. For example, the sentence “The school should reduce the tuition” includes negative reviews, as opposed to the sentence “The school reduces the tuition.”
- Intensified and diminished words: These words increase or decrease the sentimental value of the accompanying word. For example, in the phrase “She studies hard”, the element “hard” will have a stronger emotional value than in “She studies quite hard” and has a weaker emotional value than in the sentence “She studies very hard.” This relationship is caused by words of reinforcement and mitigation (e.g., very, quite) that appear alongside emotions (e.g., hard).
- Contrast phenomenon: When the text appears from the link as “though, but v.v.” an emotional valence shift occurs. Usually, the emotional tendency of the entire sentence only shows on one side. For example, “This laptop is beautiful, but the price is too expensive” is a derogatory remark because of the phrase “too expensive” due to the linking word “but” that appears earlier in the sentence.
- Emotions in the narrative: This case is not recorded because the emotion is not from the subject in question. For example, the sentence “She said that the car is very bad” where “the car is very bad” is not the opinion of the subject who made the comment. In this case, for example, the subject may add, “I think otherwise, that car is not bad at all.”
- Inconsistency (incompatibility): Many cases exist where the emotional tendency of a text depends heavily on the context where sentences and paragraphs cannot provide the actual emotional value through a calculation of the sum of the scores for each component word. Examples: (1) “The math prodigy didn’t solve that problem.” (2) “He works horribly. Every day he stayed up until 2 am to study.” (3) “BPhone is advertised with a majestic, beautiful appearance, where everything is considered to be the best, and thought to be a delicious cellphone, who would have thought it was very bad!” Similarly, for film reviews, the first or last sentences often include a dominant opinion for the entire commentary. Example “The film was released by a famous actor. The commercial was too epic. The director was too good. But it was so bad!”
Valence-shifting occurs when the semantic value of a word could be changed in a specific context [11]. Machine learning methods with bag-of-words and n-grams do not consider the effects of negation structures and other sentiment valence shifter structures, such as in the two sentences “I like this hotel but the price is quite high.” and “I like this hotel but the price is too high.” Through a bag-of-words model, the sentiment value of the two sentences is likely the same because they include the same emotional words of “like” and “high.” However, the bag-of-words model does not consider the words “quite” and “too,” which instead have the effect of making the sentiment value of the two sentences different.

Work as in [12] uses sequence mining to extract patterns of valence-shifting, such as negation, contrasting, intensifying, and diminishing of polarization. With rule-based methods and a combination of techniques, SO-CAL [13] is a pioneering system that handles valence-shifting with rules and sets of words with emotional annotations. The authors in [14] and [15] apply a dependency tree containing rich syntactic structure information to formulate syntactic rules for determining the impact of negation and other valence-shifting structures on the emotional value of sentences or the entire document.

Here, we reviewed simple but effective workarounds of using rules to capture contextual valence-shifting situations. With the development of deep learning models based on neural network methods, such as the Long short-term memory (LSTM) network [16], the attention mechanism proves to be a powerful method for modelling context, as is detailed in Section II.B.

B. ATTENTION MECHANISM
Deep learning is a research direction within the field of machine learning that has brought about results with superior efficiency creating a revolution in technology and applications of artificial intelligence in life. Deep learning has helped break through many limitations of artificial intelligence systems in various fields of study from voice processing, imaging processing, and text analysis. Moreover, deep learning has brought about a great deal of interest in the automatic learning of feature representations, such as Word2Vec [17], Glove [18]. These embedded features offer many effects and meanings to handle a variety of data types more easily from text, image, and video formats as well as creating packages to apply them to many different problems with a variety of techniques of machine learning.

The attention mechanism is an influential idea from the deep learning community and was first proposed by Bahdanau et al. in 2014 [19] for machine translation tasks. Before this approach, translation tasks relied on reading complete sentences and encoding all information into a fixed-length vector. So, a sentence with hundreds of words could be represented by only several words. This approach would lead to information loss and inadequate translations. The mechanism of attention overcomes this limitation by allowing the machine translator to scan all information in the source sentence and create the appropriate word according to the current word on which it operates as well as the context. If words are identified as being inherently important or unimportant, then models without the attention mechanism might work well because the model would automatically assign low weights to irrelevant words. However, the importance of words is highly dependent on the context. For example, the word “good” might appear in a review with the lowest rating because users are only satisfied with part of the product or service or because they use it within a negation, such as “not good”. The attention mechanism has been used successfully in parsing [20], natural language question answering [21], image question answering [22]. The application of the attention mechanism to the task of sentiment analysis has received attention in the research community, such as with document level problems by Yichun Yin and Song [23]. Also, Zhou et al. [24] used featured representations with word embeddings and the LSTM model [16] combined with a hierarchical attention mechanism. At the aspect level, the authors in work of [25] proposed the attention-based method on the sentiment classification task. The system used the two types of attentions, global and local, to classify sentence text into positive, negative, or neutral. The proposed approach can achieve higher score than that of other models.

C. HYBRID APPROACH
Studies have shown that both semantic and machine learning approaches for sentiment analysis have been applied in many different fields, each providing advantages and limitations. Recently, hybrid methods to improve the performance of these model systems have been proposed. For example, Khan et al. [26] proposed a hybrid method to predict the sentiment trends of movie reviews using SentiWordNet and machine learning. They built an emotional dictionary based on mutual information and applied a supervised learning method to detect sentiment polarization. In their research, only the scores of the adjectives in the document are considered, and the proposed algorithm results in a higher average accuracy compared to SentiWordNet. Table 1 lists notable studies that used hybrid approaches.

D. ENSEMBLE LEARNING
Combining models with different capabilities in an appropriate way can create a stronger bonding model than using individual models alone. Ensemble learning follows this idea as well as Condorcet’s theory of voting [31] that states if the probability of voting by an independent voter is \( p > 1/2 \) (i.e., each voter votes as they wish), then more voters provides increased accuracy of the majority decision. The probability of the majority vote approaches 1 as the number of voters increases.

Three methods are commonly used to combine prediction models are:

- **Bagging** [32]: Building multiple parallel operation models on different sub-training sets taken from the training...
data set. This approach can reduce overfitting in complex models.

- Boosting [33]: Building multiple models in sequential operation, such that each model learns how to correct the mistakes of the previous model in the series. The drawback of this strategy is that the training data set must be very large.

- Stacking [34]: Building multiple base-learners and a meta-learner that learns how to best combine the prediction results following those from the base-learners. This paper applies a stacking strategy for building our classification model, as illustrated in Figure 1.

By using multiple learners, the combined ability of the ensemble can be greater than a single learner [35]. Ensemble learning has been applied in many fields, such as bioinformatics [36], finance [37], and healthcare [38]. Ensemble learning applied to sentiment analysis has demonstrated advantages for many classification problems as shown by recent studies [39], [40].

### III. MATERIALS AND METHODS

#### A. DATASETS

In this study, we use three review datasets as listed in Table 2. The first dataset includes hotel reviews in Vietnam (HOTEL-Reviews) with comments posted on mytour.vn from August 2010 to July 2017. This set consists of 3,728 reviews with an average length of 47 words and separated into two positive and negative groups at a ratio of 50:50 for the training and testing sets.

The second set includes student comments about a university (UIT-VSFC) [41], and consists of 10,280 reviews with an average length of 15 words. The reviews are separated into two positive and negative groups at a ratio of 50:50 for the training and testing sets.

The third set includes food reviews in Vietnam (FOODY-Reviews) with comments on restaurants, cafes, and food posted by users on foody.vn from August 2016 to July 2019. This set consists of 40,000 reviews with an average length of 102 words and separated into two positive and negative groups at a ratio of 50:50 for the training and testing sets.

All comment data is pre-processed by eliminating misspelled words, adjusting abbreviations, social networking languages, and symbols.

#### B. PROPOSED MODEL

We adopt an ensemble learning model comprised of various base-learners to work on multiple datasets. The input training datasets are labeled as positive or negative, and the texts are standardized (correcting spelling errors, standardizing abbreviations, discarding stop-words). Each text is classified by extracting feature components into polarity shifters, including negation, contrast, inconsistency, and un-shifting. Then, TF-IDF is used for text representation to create sub-datasets that are separately processed and trained with the base learners. Other machine learning and deep learning base-learners are applied to all datasets with the word embeddings used.
FIGURE 2. The architecture of the proposed ensemble learning system.

for text representation. Finally, the base-learners’ results are merged through a meta-learner to provide a final prediction of positive or negative. Figure 2 describes the entire ensemble learning process.

1) CONTEXTUAL VALENCE-SHIFTING BASE-LEARNERS

Our classifiers are designed based on Contextual Valence-shifting as reviewed Section II.A. This study also incorporates sentiment dictionaries for determining the sentiment scores of words, as we empirically determined to increase the accuracy of the model.

a: SENTIMENT DICTIONARY

Building the sentiment dictionary is performed by the following steps.

Step 1: Use the online dictionary Vdict.com to translate text to Vietnamese from the English SentiWordnet. Here, the article translates each synset in SentiWordNet into Vietnamese, denoted by $w$, and stores its score, denoted by $SWN(w)$.

Step 2: Equate the sentiment score of the Vietnamese word from Step 1 to the score of the corresponding English word in SentiWordnet. If a word has multiple sentiment values, then select the value with the smallest difference between the word’s score obtained by the results of the training process using logistic regression. The process for determining the sentiment score of the vocabulary set with logistic regression is as follows.

The sentiment score of each word (denoted as $LR(w)$) is determined by the pseudocode:

for $w$, $i$ in word_index_map
  $LR(w) = model.coef[i]$
  if $|LR(w)| > \text{threshold}$ then return $w$ and $LR(w)$

The complete algorithm is illustrated as follows.

Algorithm VNSD.[43]

Input: list of $LR(w)$
Output: list of $w_{\text{final}}$
Method:

- $LR(w)$: score of the lexical item $w$ is extracted from the Logistic Regression (LR) training task.
- $w_i$: score of the $i^{th}$ of lexical item $w$ is extracted from the SentiWordnet (SWN) translation task.

The final score of the lexical item $w$ is determined by the function:

function finalScore($w$)
  $w_{LR} = LR(w)$
  for each $w_i$ in the set of $SWN(w)$:
    return $w_i$
  if $w_i$ makes $|w_i - w_{LR}|$ get a minimum value

Algorithm VNSD results in a set of sentiment words and a corresponding score that is separated into lists of positive and negative words sorted in descending order by the absolute values of the words.

b: CONTEXTUAL VALENCE-SHIFTING BASE-LEARNERS

We adopted Logistic Regression (LR), n-gram representation and bag-of-words TF-IDF. The training process for the entire dataset and the sub-datasets of negation, contrast, inconsistency, and un-shifting is as follows.

Negation Sub-Dataset: Negation structure is the most common form of valence-shifting. Determining the negation structure is performed by checking for the appearance of negation words, such as “no,” “not,” “none,” “don’t,” and “never” in a sentence after removing exceptions, such as “not only . . . but also . . .”, which cannot be considered as a negation.

These reviews are added to the negation dataset, and after determining the location of the negation word, it is removed and a negative sentiment word replaces the first sentiment word behind the negation word according to the score in the sentiment dictionary. The sentiment words behind this first word will also be replaced if they are identified to have the same sentiment polarity as this word. For example: (1)
"I don’t like this hotel!” will be changed to “I dislike this hotel!” (2) “The staff is not friendly, enthusiastic.” will be changed to “The staff is unsympathetic, unenthusiastic.”

**Contrast Sub-Dataset:** Contrast structures are also common for valence-shifting. These words are separated into two groups of Fore-Contrast, such as “but” and “however”, and Post-Contrast, such as “although” and “although”. If the sentence contains a Fore-Contrast word, then the valence shift occurs in the clause immediately preceding this word. If the sentence contains a Post-Contrast word, then the valence shift occurs within the clause, in which the Post-Contrast word appears. These contrast sentences are then added to the contrast dataset. For example, “The hotel is very nice, convenient location but the price is a bit expensive” results in a valence shift of “The hotel is very nice, convenient location but the price is a bit expensive”.

**Inconsistency Sub-Dataset:** Inconsistency sentences do not show signs of a grammatical shift, but the semantics provides an opposite sentiment to the sentiment expressed in the entire text. This conflict is due to the complexity of human language. Sentimental inconsistency can also be identified by scoring each word in the text, and the review can be estimated by a valence shift with

\[
h(r_i) = y \cdot \sum_{j=0}^{|r_i|} w(t_j)
\]

where \(y\) is the positive or negative label of the review, \(r_i\) is the \(i\)th review, \(|r_i|\) is the number of words in the \(i\)th review, and \(w(t_j)\) is the sentiment score of the word \(t_j\).

Based on the estimated value, one of the following decisions is made:

- If \(h(s_i) < 0\), then the review is added to the inconsistency set containing the inconsistency sentiment reviews.
- If \(h(s_i) \geq 0\), then the review is added to the un-shifting set containing the un-shifted sentiment reviews.

Examples of sentiment inconsistencies are:
1. “More specifically, he has spent a lot of effort.”
2. “The teacher works so horribly every day he stays up until 2 am.”

**Un-Shifting Sub-Dataset:**

Datasets containing sentences without valence-shifting are considered un-shifted. In other words, the un-shifted dataset includes the text that remains after the removal of sentences that contain a valence-shifting cases of negation, contrast, or inconsistency.

2) **ATTENTION BASE-LEARNER**

We apply the bidirectional long short term memory (Bi-LSTM) network [16] with an attention mechanism to design an additional base-learner. Similar to Bi-GRU [44], this model leverages deep learning to capture the contextual factors in a text. In our experiments, Word2Vec word embedding for all datasets initializes using 300-dimensional word vectors pre-trained by [45].

The architecture of the attention mechanism model is illustrated in Figure 3. First, the model maps words into low-dimensional word vectors using pre-trained word embeddings from the input layer. Given \(S = \{w_1, w_2, w_3, \ldots, w_n\}\), each word in \(S\) is mapped into a \(k\)-dimensional vector \(e_i \in \mathbb{R}^k\) through a lookup from a pre-trained word embedding matrix \(E \in \mathbb{R}^{k \times |V|}\), such as Glove [18] and Word2vec, such that

\[
e_i = E(w_i)
\]

where \(k\) is the dimension of the word vector and \(|V|\) is the vocabulary size.

Next, the Bi-LSTM operates on these word vectors to preserve the sequential information in a memory modelling layer. The forward LSTM learns the forward weight of the hidden state to obtain the \(i\)th forward context vector \(\vec{c}_i\). Given the hidden state \(\vec{c}_{t-1} \in \mathbb{R}^d\) and the word embedding \(e_i\), the hidden state \(\vec{c}_t\) at time step \(t\)th can be calculated as

\[
\vec{c}_t = LSTM(e_i, \vec{c}_{t-1})
\]

The backward LSTM learns the backward weight of the hidden state to obtain the \(i\)th backward context vector \(\vec{c}_i\). Then, a concatenation of \(\vec{c}_t\) and \(\vec{c}_i\) provides the \(i\)th context vector \(c_i \in \mathbb{R}^{2d}\) of each word as

\[
c_i = \tanh(\vec{c}_t; \vec{c}_i)
\]

where [;] stands for the concatenate operation, and \(\vec{c}_t\) and \(\vec{c}_i\) are the output of the forward and backward LSTM at time step \(t\)th, respectively. We next use a word-level context vector \(u_w\) that is randomly initialized to measure the importance of the word as the similarity of \(u_t\). The context vector \(u_w\) is a high-level representation of the informative word, and the value...
of $u_w$ is updated during the training process. We obtain the attention score $\alpha_t$ normalized by the $\text{softmax}$ function.

$$u_t = \tanh(W_w c_t + b_w)$$ (5)

$$\alpha_t = \text{softmax}(u_t, u_w)$$ (6)

After this process, we compute the document vector as a weighted sum of the word annotations at the current time in the decoded state using

$$s = \sum(\alpha_t, c_t)$$ (7)

Finally, in a $\text{softmax}$ layer, the sentiment polarity distribution predicts the probability $p_b$ of the text belonging to the category $k$ (positive or negative), such that

$$p = \text{softmax}(W_s s + b_s)$$ (8)

where $W_s$ represent weight matrices, and $b_s$ are bias vectors.

3) META-LEARNER
A meta-learner is trained from the combined predictions of the previous base-leaners. These base predictions are input into a sequential layer and combined to form a new set of predictions. In this study, the different base-learners are fit independently, and the meta-learner is trained on top of these to predict an output based on the outputs of the base-learners. Different algorithms can be selected for doing this part of the sentiment classification, and we apply logistic regression for our meta-learner as it suggested the best results through experimentation.

IV. EMPIRICAL EXPERIMENTS
A. MODEL COMPARISONS
To verify the effectiveness of the proposed model, we tested a series of variations of the model. First, five base-learners were used, ignoring the deep learning base-learner. Tests were conducted with and without the use of sentiment dictionaries. Next, we estimated the model using six base-learners and included the attention-based model with and without sentiment dictionaries. Finally, we attempted to compare our method with other baseline models of SVM, LSTM, BiLSTM [46], attention-based, BERT-based [47], and the state-of-the-art model CEM in [48].

- SVM-based model: The traditional machine learning method of SVM representing the TF-IDF score for different unigrams.
- LSTM-based model: The Long Short-Term Memory (LSTM) Network proposed by Hochreiter and Schmidhuber [16] as enhanced by Gers [49]. Numerous deep learning models developed from LSTM have been proposed but the original LSTM model remains valuable as a strong baseline model [50]. We use one LSTM network to model the context, which includes 2 hidden layers, 64-units, and Word2Vec feature representation with a one-hot vector dimension of 65,000, reduced to 300 after applying Word Embedding. The average value of all hidden states is treated as the final context representation.
- BiLSTM-based model: We use bidirectional LSTM networks model to model the context, which includes 128-units, and Word2Vec feature representation with a one-hot vector dimension of 65,000, reduced to 300 after applying Word Embedding.
- Attention-based model: A vanilla global attention mechanism applied to the output of BiLSTM. To mine deeper feature information from the text, a stacked hierarchical training mechanism is used to train the data, and multi-layer feature learning obtains a more representative feature representation.
- BERT-based model: The Bidirectional Encoder Representations from Transformers (BERT) is a state-of-art language model that produces contextualized representation learning [47], and within this model, the Masked Language Modeling (MLM) task embeds each word based on surrounding words. The Next Sentence Prediction (NSP) task, on the other hand, learns semantic coherence between sentences. The BERT model archives positive results in multiple natural language processing tasks, including sentiment analysis [51]. We use the fine-tuned BERT-based model proposed by Nguyen\(^1\) that outperforms the result of the winning model from the AIViVN’s sentiment classification contest.\(^2\)

The proposed model and its variants include the following.

- CEM(6C-LSTM-WLLR) model: As proposed in [48], this six base-learners approach includes a contrast learner, inconsistency learner, negation learner, un-shifting learner, logistic regression learner, and LSTM learner with a meta-learner using Logistic Regression.
- CEM(6C-ATT-WLLR) model: This model includes six base-learners likes the CEM(6C-LSTM-WLLR) as in [48], but the LSTM model is replaced with the attention mechanism. The base-learners include a contrast learner, an inconsistency learner, a negation learner, an un-shifting learner, a logistic regression learner, an attention learner, and the meta-learner which uses Logistic Regression.
- CEM(6C-ATT-VNSD) model: This model includes six base-learners likes the CEM(6C-LSTM-WLLR) as in [48], but the LSTM model is replaced with the attention mechanism and the WLLR (weighted log-likelihood ratio) statistical method as in [48] is replaced with the sentiment dictionaries (VNSD) to consider for sentiment score ranking. The base-learners include a contrast learner, inconsistency learner, negation learner, un-shifting learner, logistic regression learner, and attention learner, and the meta-learner uses Logistic Regression.

---

\(^1\)https://github.com/suicaophoBERT-Sentiment-Classification

\(^2\)https://www.aivivn.com/contests/6
TABLE 3. Experimental results on the HOTEL-Reviews dataset with a meta-learner using logistic regression.

| Model                        | HOTEL-Reviews |
|------------------------------|---------------|
| SVM-based                    | 94.15%        |
| LSTM-based                   | 92.75%        |
| BiLSTM-based                 | 93.59%        |
| Attention-based              | 93.29%        |
| BERT-based                   | 89.84%        |
| CEM(5C-LSTM-WLLR) [48]       | 93.07%        |
| CEM(6C-LSTM-WLLR) [48]       | 94.36%        |
| CEM(6C-ATT-WLLR)             | 96.03%        |
| CEM(6C-ATT-VNSD)             | 93.61%        |
| CEM(6C-ATT-VNSD)             | 94.63%        |

B. EVALUATION RESULTS
The experimental results are evaluated with the accuracy ratio, as shown in Tables 3, 4, and 5.

1) EVALUATION RESULTS FOR HOTEL-REVIEWS
HOTEL-Reviews is a small dataset which has a lot of social networking slangs. The deep learning models that need rich training data such as BERT-based, LSTM, BiLSTM, and the attention mechanism, which give a lower accuracy than traditional machine learning methods like SVM. The models proposed in this article all have good performance – CEM(6C-ATT-WLLR) showed the best performance and CEM(6C-ATT-VNSD) was the second-best. Based on the experimentation in this study, the proposed models are better than the best model in [48] by 0.3%–1.7%, which is CEM(6C-LSTM-WLLR).

2) EVALUATION RESULTS FOR UIT-VSFC
UIT-VSFC is a medium-sized dataset with well-written comments. Through experiments, we learn that the deep learning models, such as BERT-based, BiLSTM, LSTM, and the attention mechanism give superior results in comparison to traditional machine learning methods such as SVM. The ensemble learning model was superior with sufficiently large training data when CEM(6C-ATT-WLLR) and CEM(6C-ATT-VNSD) performed well. The proposed models are better than the best model in [48] – CEM(6C-LSTM-WLLR) – by up to 1.65%.

3) EVALUATION RESULTS FOR FOODY-REVIEWS
FOODY-Reviews is a medium-large-sized dataset and with comments written on social networks. Through these experiments, we learn that traditional machine learning methods, such as SVM offer more accurate results compared to deep learning models, such as LSTM, BiLSTM, and the attention mechanism, except for the BERT-based model, which works well with relatively sufficient data.

The ensemble learning model was superior when sufficient training data was available and had good performance with CEM(6C-LSTM-WLLR), CEM(6C-ATT-WLLR), and CEM(6C-ATT-VNSD). The proposed models are superior to CEM(6C-LSTM-WLLR) – the best model in [48] – by up to 1.89%.

C. SYSTEM ANALYSIS
Some important observations made through the experiments on three datasets are:

• In addition to the ensemble learning model, a deep learning base-learner allows the model to achieve better results than the model with five base-learners. Tables 3, 4, and 5 show a comparison between CEM(5C-WLLR) and CEM(6C-LSTM-WLLR), CEM(6C-ATT-WLLR), and CEM(6C-ATT-VNSD), respectively. This demonstrates the automatic learning features of deep learning that contribute to the improvement of system performance.

• The use of the attention mechanism in CEM(6C-ATT-WLLR), instead of LSTM in CEM(6C-LSTM-WLLR), helps ensemble learning systems obtain superior classification results; however, the attention mechanism itself (Attention-based) does not appear to be superior to LSTM in sentiment classification, as shown in Tables 3, 4, and 5.

• The use of sentiment dictionaries VNSD in CEM(6C-ATT-VNSD), instead of WLLR in CEM(6C-LSTM-WLLR) [48], helps ensemble-learning systems achieve better classification results, as shown in Tables 3, 4, and 5. Sentiment dictionaries can obtain
more accurate emotional scores than the WLLR statistical method, which enhances the performance of the system.

- The BERT-based model also outperforms other deep learning models that require larger amounts of training data, as listed in Tables 4 and 5. The opposite is not good for small data (HOTEL-Reviews, Table 3).

- When sufficient training data is provided, as shown in Table 5, the BERT-based model results are nearly equivalent to the proposed model (CEM).

- With comments written on social networks, we learn that traditional machine learning method SVM gives more accurate results than deep learning models such as LSTM, BiLSTM, and Attention-based, as shown in Tables 3 and 5.

- In Table 5, the ensemble learning model with the use of sentiment dictionaries CEM(6C-ATT-VNSD) was superior when sufficient training data was available and the reviews are quite long. On the contrary, CEM(6C-ATT-WLLR) achieves better results than CEM(6C-ATT-VNSD), as shown in Tables 3 and 4.

In this section, we also carefully analysed both the correctly and incorrectly predicted samples in the three datasets. Most errors occur when opinions are expressed in an implicit way. For example, in the review (1) “Tôi nâng cấp phần ăn của mình lên 50k.” (English: “I upgraded my meal to 50k.”) contains facts without any opinion words, as more positive samples in the training may bias the neutral sentence to seem positive. In this case, the sentiment score cannot be determined by detecting explicit opinion words, but rather require a deep understanding of the entire sentence. Gaining a better understanding of implicit opinions remains challenging in deep learning methods. With the reviews (2) and (3), the system predicted correctly (as negative).

- Review (2) “Bảng nữa giá học phí so với Huflit thì tốt.” (English: “It would be great if the tuition fee is half as much as Huflit’s.”) contains facts without any opinion words, as more positive samples in the training may bias the neutral sentence to seem positive. In this case, the sentiment score cannot be determined by detecting explicit opinion words, but rather require a deep understanding of the entire sentence. Gaining a better understanding of implicit opinions remains challenging in deep learning methods. With the reviews (2) and (3), the system predicted correctly (as negative).

In Figure 4, the ensemble learning model with the use of sentiment dictionaries CEM(6C-ATT-VNSD) was superior when sufficient training data was available and the reviews are quite long. On the contrary, CEM(6C-ATT-WLLR) achieves better results than CEM(6C-ATT-VNSD), as shown in Tables 3 and 4.

V. CONCLUSIONS

This paper introduced an effective ensemble learning model for sentiment classification. Our system can adaptively capture contextual information in text reviews by combining rule-based methods and other state-of-the-art deep learning models. We adopted word embedding representation and the attention mechanism, along with pre-defined rules and specific-domain sentiment dictionaries, to identify numerous valence-shifting cases. Although the computational cost of the proposed system is higher than those of compared algorithms, the model has multiple distinctive characteristics and gives better results than other approaches.

In the future, the following related tasks will be considered. First, we will attempt to improve the performance of this model by using other techniques like BERT-based, ELMo-based [52]. Second, an online application will be introduced to reuse of these results.

ACKNOWLEDGMENT

The authors would like to thank Ho Chi Minh City University of Technology (HCMUT), and VNU-HCM for their support.

REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: Sentiment classification using machine learning techniques,” in Proc. ACL Conf. Empirical Methods Natural Lang. Process. (EMNLP), vol. 10, 2002, pp. 79–86, doi: 10.3115/1118693.1118704.

[2] P. D. Turney, “Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews,” in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2001, p. 417, doi: 10.3115/1073083.1073153.

[3] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Found. Trends Inf. Retr., vol. 2, nos. 1–2, pp. 1–135, 2008.

[4] C. Fellbaum, WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press, 1998.

[5] B. Stefano, E. Andrea, and F. Sebastiani, “SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining,” in Proc. 7th Int. Conf. Lang. Resour. Eval. (LREC), vol. 10, 2010, pp. 2200–2204.

[6] Y. Cheng, “Agreement-based joint training for bidirectional attention-based neural machine translation,” in Joint Training for Neural Machine Translation. Springer, 2019, pp. 11–23.

[7] H. Zhang, J. Cai, J. Xu, and J. Wang, “Pretraining-based natural language generation for text summarization,” in Proc. 23rd Conf. Comput. Natural Lang. Learn., 2019, pp. 789–797.

[8] X. Guo, H. Zhang, H. Yang, L. Xu, and Z. Ye, “A single attention-based combination of CNN and RNN for relation classification,” IEEE Access, vol. 7, pp. 12467–12475, 2019, doi: 10.1109/ACCESS.2019.2891770.

[9] F. Long, K. Zhou, and W. Ou, “Sentiment analysis of text based on bidirectional LSTM with multi-head attention,” IEEE Access, vol. 7, pp. 141960–141969, 2019, doi: 10.1109/ACCESS.2019.2942614.

[10] S. Rida-E-Fatima, A. Javed, A. Banjar, A. Iriz, H. Dawood, H. Dawood, and A. Alamri, “A multi-layer dual attention deep learning model with refined word embeddings for aspect-based sentiment analysis,” IEEE Access, vol. 7, pp. 114795–114807, Aug. 2019, doi: 10.1109/Access.2019.2927281.

[11] L. Polanyi and A. Zemken, “Contextual Valence Shifters,” in Computing Attitude and Affect in Text: Theory and Applications. Berlin, Germany: Springer-Verlag, 2006, pp. 1–10.

[12] G. Xu and C.-R. Huang, “Extracting Chinese polarity shifting patterns from massive text corpora,” Lingua Sinica, vol. 2, no. 1, p. 5, Dec. 2016, doi: 10.1186/s40655-016-0014-z.

[13] M. Taboada, J. Brooke, M. Tofilloski, K. Voll, and M. Stede, “Lexicon-based methods for sentiment analysis,” Comput. Linguistics, vol. 37, no. 2, pp. 267–307, Jun. 2011, doi: 10.1162/coli_a_00049.

[14] I. Carrillo de Albornoz, L. Plaza, and P. Gervás, “A hybrid approach to emotional sentence polarity and intensity classification,” in Proc. 14th Conf. Comput. Nat. Lang. Learn., 2010, pp. 153–161.

[15] L. Jia, C. Yu, and W. Meng, “The effect of negation on sentiment analysis and retrieval effectiveness,” in Proc. 18th ACM Conf. Inf. Knowl. Manage. (CIKM), 2009, p. 1827, doi: 10.1145/1645953.1646241.

[16] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
T. K. Tran, T. T. Phan: Capturing Contextual Factors in Sentiment Classification: An Ensemble Approach

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in Proc. Int. Conf. Learn. Represent., 2013.

[18] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2014, pp. 1532–1543, doi: 10.3115/v1/d14-1162.

[19] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in Proc. Int. Conf. Learn. Represent., 2014.

[20] T. Dozat and C. D. Manning, “Deep biasfree attention for neural dependeneparing,” in Proc. 5th Int. Conf. Learn. Represent. (ICLR), 2016.

[21] X. He and D. Golub, “Character-level question answering with attention,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2016, pp. 1598–1607, doi: 10.18653/v1/d16-1166.

[22] J. Liu, J. Yang, D. Batra, and D. Parikh, “Hierarchical question-image co-attention for visual question answering,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2016, pp. 289–297.

[23] Y. Yin, Y. Song, and M. Zhang, “Document-level multi-aspect sentiment classification as machine comprehension,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2017, pp. 2044–2054.

[24] X. Zhou, X. Wan, and J. Xiao, “Attention-based LSTM network for cross-lingual sentiment classification,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2016, pp. 247–256, doi: 10.18653/v1/d16-1024.

[25] X. Wang, G. Xu, J. Zhang, X. Sun, L. Wang, and T. Huang, “Syntactic-directed hybrid attention network for aspect-level sentiment analysis,” IEEE Access, vol. 7, pp. 5014–5025, 2019, doi: 10.1109/ACCESS.2018.2850352.

[26] F. H. Khan, U. Qamar, and S. Bashir, “A semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet,” Knowl. Inf. Syst., vol. 51, no. 3, pp. 851–872, Jun. 2017.

[27] Y. Ding, J. Yu, and J. Jiang, “Recurrent neural networks with auxiliary labels for cross-domain opinion target extraction,” in Proc. 31st AAAI Conf. Artificial Intell., 2017, pp. 3436–3442.

[28] P. Chikkerur, S. Porta, and E. Cambria, “SENUT: Sentiment analysis of tweets by combining a rule-based classifier with supervised learning,” in Proc. 9th Int. Workshop Semantic Eval. (SemEval ), 2015, pp. 647–651, doi: 10.18653/v1/s15-2108.

[29] R. Ortega, A. Fonseca, Y. Gutiérrez, and A. Montoyo, “SSA-UO: Unsupervised Twitter sentiment analysis,” in Proc. Int. Workshop Sem-Mantic Eval. (SemEval), 2013.

[30] P. P. B. Filho, L. Avanço, T. A. S. Pardo, and M. G. V. Nunes, “NILC-USP: An improved hybrid system for sentiment analysis in Twitter messages,” in Proc. 5th Int. Workshop Semantic Eval. (SemEval), vol. 2, 2014, pp. 568–572, doi: 10.3115/v1/a4-1074.

[31] H. P. Young, “Condorcet’s theory of voting,” Amer. Political Sci. Rev., vol. 82, no. 4, pp. 1231–1244, Dec. 1988, doi: 10.2307/1967157.

[32] L. Breiman, “Bagging predictors,” Mach. Learn., vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1007/bf00058655.

[33] Y. Freund and R. E. Schapire, A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting (Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 904, 1995, pp. 23–37, doi: 10.1007/3-540-59119_2_166.

[34] L. Breiman, “Stacked regressions,” Mach. Learn., vol. 24, no. 1, pp. 49–64, 1996, doi: 10.1007/bf00117832.

[35] P. Domingos, “A few useful things to know about machine learning.” Commun. ACM, vol. 55, no. 10, pp. 78, 2012, doi: 10.1145/2347736.2347755.

[36] A. Verma and S. Mehta, “A comparative study of ensemble learning methods for classification in bioinformatics,” in Proc. 7th Int. Conf. Cloud Comput., Data Sci. Eng. Confluence, Jan. 2017, pp. 155–158, doi: 10.1109/CONFLUENCE.2017.7943141.

[37] G. Xie, Y. Zhao, M. Jiang, and N. Zhang, “A novel ensemble learning approach for corporate financial distress forecasting in fashion and textiles supply chains,” Math. Problems Eng., vol. 2013, pp. 1–9, Mar. 2013, doi: 10.1155/2013/493931.

[38] Y. Li, C. Bai, and C. K. Reddy, “A distributed ensemble approach for mining healthcare data under privacy constraints,” Inf. Sci., vol. 330, pp. 245–259, Feb. 2016, doi: 10.1016/j.ins.2015.10.011.

[39] O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. A. Iglesias, “Enhancing deep learning sentiment analysis with ensemble techniques in social applications,” Expert Syst. Appl., vol. 77, pp. 236–246, Jul. 2017, doi: 10.1016/j.eswa.2017.02.002.

[40] R. Kumar, P. Li, J. Yu, Y. Qi, and E. Cambria, “Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis,” Inf. Process. Manage., vol. 52, no. 1, pp. 36–45, Jan. 2016, doi: 10.1016/j.ipm.2015.04.003.

[41] K. V. Nguyen, V. D. Nguyen, P. X. V. Nguyen, T. T. H. Truong, and N. L.-T. Nguyen, “UIT-VSFC: Vietnamese Students’ feedback corpus for sentiment analysis,” in Proc. 10th Int. Conf. Knowl. Syst. Eng. (KSE), Nov. 2018, pp. 19–24, doi: 10.1109/KSE.2018.8573337.

[42] W. Frank, S. L. Koppen, L. G. M. Noordman, and W. W. Monk, Modeling Multiple Levels of Text Representation. Mahwah, NJ, USA: Edfbau, 2007.

[43] T. K. Tran and T. T. Phan, “A hybrid approach for building a vietnamese sentiment dictionary,” J. Intell. Fuzzy Syst., vol. 35, no. 1, pp. 967–978, Jul. 2018, doi: 10.3233/JIFS-172053.

[44] Y. Han, M. Liu, and W. Jing, “Aspect-level drug reviews sentiment analysis based on double BiGRU and knowledge transfer,” IEEE Access, vol. 8, pp. 21314–21325, 2020, doi: 10.1109/ACCESS.2020.2969473.

[45] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, and M. Johnson, “VnCoreNLP: A vietnamese natural language processing toolkit,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Demonstrations, 2018, pp. 58–60, doi: 10.1109/ACCESS.2019.2909919.

[46] L. K. J. Devlin and M. W. Chang, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Human Lang. Technol., 2019, pp. 4171–4186.

[47] K. Tran and T. Phan, “Deep learning application to ensemble learning—The simple, but effective, approach to sentiment classifying,” Appl. Sci., vol. 9, no. 13, pp. 2760, Jul. 2019, doi: 10.3390/app9132760.

[48] G. Melis, C. Dyer, and P. Blunsom, “On the state of the art of evaluation in neural language models,” in Proc. 6th Int. Conf. Learn. Represent. (ICLR), 2018.

[49] X. Li, X. Fu, G. Xu, Y. Yang, J. Wang, L. Jin, Q. Liu, and T. Xiang, “Enhancing BERT representation with context-aware embedding for aspect-based sentiment analysis,” IEEE Access, vol. 8, pp. 46686–46687, 2020, doi: 10.1109/ACCESS.2020.2978511.

[50] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and J. Zettlemoyer, “Deep contextualized word representations,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol., vol. 1, 2018, pp. 2227–2237, doi: 10.1109/ACCESS.2018.2817632.