Article

Deep Learning Application to Ensemble Learning—The Simple, but Effective, Approach to Sentiment Classifying

Thien Khai Tran [1,2,*] and **Tuoi Thi Phan [1]**

[1] Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology, VNU-HCM, Ho Chi Minh City, Vietnam
[2] Faculty of Information Technology, Ho Chi Minh City University of Foreign Languages and Information Technology (HUFLIT), Ho Chi Minh City, Vietnam
* Correspondence: thientk@cse.hcmut.edu.vn or 8141216@hcmut.edu.vn or thientk@huflit.edu.vn

check for updates

Abstract: Sentiment analysis is an active research area in natural language processing. The task aims at identifying, extracting, and classifying sentiments from user texts in post blogs, product reviews, or social networks. In this paper, the ensemble learning model of sentiment classification is presented, also known as CEM (classifier ensemble model). The model contains various data feature types, including language features, sentiment shifting, and statistical techniques. A deep learning model is adopted with word embedding representation to address explicit, implicit, and abstract sentiment factors in textual data. The experiments conducted based on different real datasets found that our sentiment classification system is better than traditional machine learning techniques, such as Support Vector Machines and other ensemble learning systems, as well as the deep learning model, Long Short-Term Memory network, which has shown state-of-the-art results for sentiment analysis in almost corpuses. Our model's distinguishing point consists in its effective application to different languages and different domains.

Keywords: sentiment analysis; ensemble learning; deep learning; CEM; deep features; surface features; valence shifters

1. Introduction

Sentiment classification or opinion mining is the narrow field of natural language processing, information querying, and text mining that is used to extract a person's impressions of or thoughts about something from nonstructural text data. This research domain has drawn the interest of not only scientists but also businesses and organizations worldwide. The ability to classify sentiment has a practically tremendous impact practically because it helps businesses save the expenditure of human resources to determine customers' needs while helping customers choose more suitable products and services according to their necessities. Li and Liu's [1] survey reported that more than 80% of internauts search at least once for reviews about a product they intend to buy before making their decision.

The problem of sentiment classification was raised by Dave et al. [2] and Nasukawa and Yi [3] in the early 2000s. Since then, many research studies have been conducted to classify and evaluate reviews about products and services in media and blog posts, which can be classified into three levels of interest: (i) document level; (ii) sentence level; and (iii) aspect level. On the first level, Tang et al.'s model [4], based on a deep learning approach, and Xia et al.'s ensemble learning model [5] should be mentioned. With reference to the sentence level, Marcheggiani et al. [6] and Yang and Cardie [7] proposed models based on conditional random fields (CRFs). However, the aspect level has received the

most consideration within various publications. The focus of this paper is on Chinsha and Joseph's [8] and Tran et al.'s [9] work, which proposed a syntactic-based approach using dependency grammar.

Techniques used for classifying sentiments can be put into three main groups: machine learning comprising Pang et al. [10], Riaz et al. [11], and Wang et al. [12]; relying on vocabulary with Turney et al. [13], Muhammad et al. [14], and Khan et al. [15]; and a hybrid of machine learning and vocabulary with Balahur et al. [16] and Keshavarz and Abadeh [17]. Machine learning techniques consist of supervised machine learning like Severyn et al. [18], semi-supervised machine learning like Hajmohammadi et al. [19], and unsupervised machine learning like Claypo and Jaiyen [20]. Techniques that rely on vocabulary comprise three groups referring to some noticeable works as follows: a vocabulary approach with Saif et al. [21], a dataset approach with Vulic et al. [22], and integration of the above ones with Taboada et al. [23].

In the last two decades, machine learning methods have dominated majority of sentiment analysis tasks. Since feature representation greatly influences the performance of a machine learning algorithm [24], many researches focus on getting effective features in-hand with domain expertise and ad hoc techniques. However, this work can be completed using representation learning algorithms, such as the deep learning approach, which automatically distinguishes and explains text representations from data. Deep learning has emerged due to its ability to represent data at various classified levels. Wu et al. [25] and Zhao et al. [26] are two remarkable evidences of this approach.

In regard to valence shifters in sentiment analysis, users often produce reviews about subjects based on the sentiment levels. Thus, the sentiment value of a phrase can be affected by the corresponding context, which is called polarity shifting (or valence shifters) [27]. Polarity shifting bears complex language structures which consist of negative, contrasting, intensified, and diminished structures [23]. Polarity shifting can make traditional approaches, such as machine learning with the Bag-of-Words (BoW) model, ineffective because these approaches are interested in whether single words bear positive or negative polarity, based on a predetermined sentiment dataset. By contrast, the common technique for sentiment classification focuses on polarity shifting, which requires an analysis of a phrase's structure and semantics [28–31].

Xia et al. [5] used four classifiers (two baseline machine learning classifiers and two statistical classifiers) with four sub-datasets, which were comprised of various valence shifting structures. Oscar et al. [32] proposed an ensemble of sentiment classifiers, where several baseline classifiers trained with different types of features were combined. The authors adopted deep learning to produce features for the classifiers automatically.

This paper adopts ensemble learning to classify sentiment at the document level inspired by [5,32]. We extract various different features from datasets for base learners by identifying the various structures that cause polarity shifting in the text; we call these 'surface features'. We also use word embedding and deep learning to extract other features, which are called 'deep features'. The proposed system was built and experiments were conducted using datasets to check the system's performance in Vietnamese and English. A comparison with other machine learning approaches showed that the results of the proposed system were better than even state-of-the-art deep learning models and other ensemble learning systems. The experimental results also show that taking 'deep features' into consideration for base learners improves the effectiveness of the system.

The following are the contributions of this research:

- We propose an effective ensemble learning system using datasets of base learners, which comprise features that result from exploring language characteristics and applying a deep learning model.
- We adopt word embedding and develop a deep learning model for base learners that helps improve the system's effectiveness.
- The proposed model proved appropriate for the Vietnamese language, and also yielded adequate results for the various English datasets.

The remainder of the paper is organized as follows: Section 2 presents related existing work, Section 3 presents the proposal of our model, Section 4 describes the experiments and valuations, and Section 5 presents the conclusion and introduces directions for future research.

2. Related Work

Polarity shifting occurs when a phrase's sentiment value changes according to a specific context [27]. The first machine learning methods failed to take account of influences caused by negation structures and other polarity shifting structures. For instance, for early machine learning, the two sentences 'The hotel is very nice but the price is high.' and 'The hotel is very nice, the price is high.' are likely to be classified into the same stage because they contain the same words indicating the sentiment 'nice' and 'high'. To overcome this issue, the works [33] have recently used sequence mining to extract polarity shifting models that inverse, decrease, and eliminate polarity. Using a hybrid of different techniques, SO-CAL (Semantic Orientation CALculator) [23] was one of the first systems to process polarity shifting using rule models and sentiment vocabulary labeled as sentiment datasets, and [34,35] used dependency grammar to define syntactic rules that identified each negative structure's influence and other polarity shifting structures.

Ensemble learning is a strong machine learning model that is optimal in classification problems involving many learners; the ability of an ensemble learning model to generalize is much better than that of a single learning model [36]. Ensemble learning is applicable in various domains, including bioinformatics [37], finance [38], and healthcare [39]. The latest research indicates that ensemble learning models could be applied to the sentiment classification problem. Table 1 shows the relevant works conducted over the last ten years that have applied ensemble learning to sentiment classification.

Table 1. Noticeable researches connected to the application of ensemble learning to sentiment classification.

Works	Features	Learning Models	Languages
Xia et al. [40]	n-gram	Naïve Bayes, Maximum Entropy and Support Vector Machines	English
Li et al. [41]	n-gram, lexicon	Support Vector Machines, k-Nearest Neighbors, Scoring	Chinese
Su et al. [42]	n-gram	Naïve Bayes, k-Nearest Neighbors, Maximum Entropy, Support Vector Machines	Chinese
Shoushan et al. [43]	n-gram	Support Vector Machines, Logistic Regression	English
Xia et al. [5]	n-gram	Support Vector Machines, Logistic Regression	English
Oscar et al. [32]	n-gram lexicon, word vectors	Scoring, Naive Bayes, Maximum Entropy, Support Vector Machines	English

Apropos the deep learning use for sentiment classification, this approach has recently been recognized as a strong machine learning model and has produced advanced results in various domains to which it has been applied—from computer vision and speech processing to natural language processing [44]. The application of deep learning to sentiment classification has also become more popular. Some recent research studies on sentiment classification that engage in maneuvering deep learning are depicted in Table 2.

Table 2. Examples of new researches on sentiment classification using deep learning.

Works	Features	Models	Languages	Level
Tang et al. [4]	Word embeddings	CNN/LSTM + GRU	English	Document/Sentence
Johnson et al. [45]	Bag of words	Bag of words CNN + Sequence CNN	English	Document
Li et al. [46]	Bag of words	CNN	Chinese	Aspect
Zhou et al. [47]	Word embeddings	LSTM	English	Document
Nguyen et al. [48]	Word embeddings + Ontology	CNN + LSTM	Vietnamese	Aspect

3. Proposed Models

3.1. Architecture

The system's input is training datasets comprising of labeled texts classified into positive and negative stages. These texts are made to pass through a preprocessing component in order to be standardized (correcting spelling errors, abbreviations, discarding stop words) and analyzed into appropriate sentences or clauses. After the preprocessing stage, every text's sentence set will be classified by feature extracting components, into polarity shifters, which are negation, contrast, inconsistency, and no_shift sentences. These sub-datasets will be separately processed to be trained with base learners. At a certain time, another base learner will be applied to all the datasets. These datasets will also be used for training with a deep learning model and word embedding representation Word2Vec [49]. Finally, the base learners' results will be integrated through ensemble learning. Figure 1 describes the whole process.
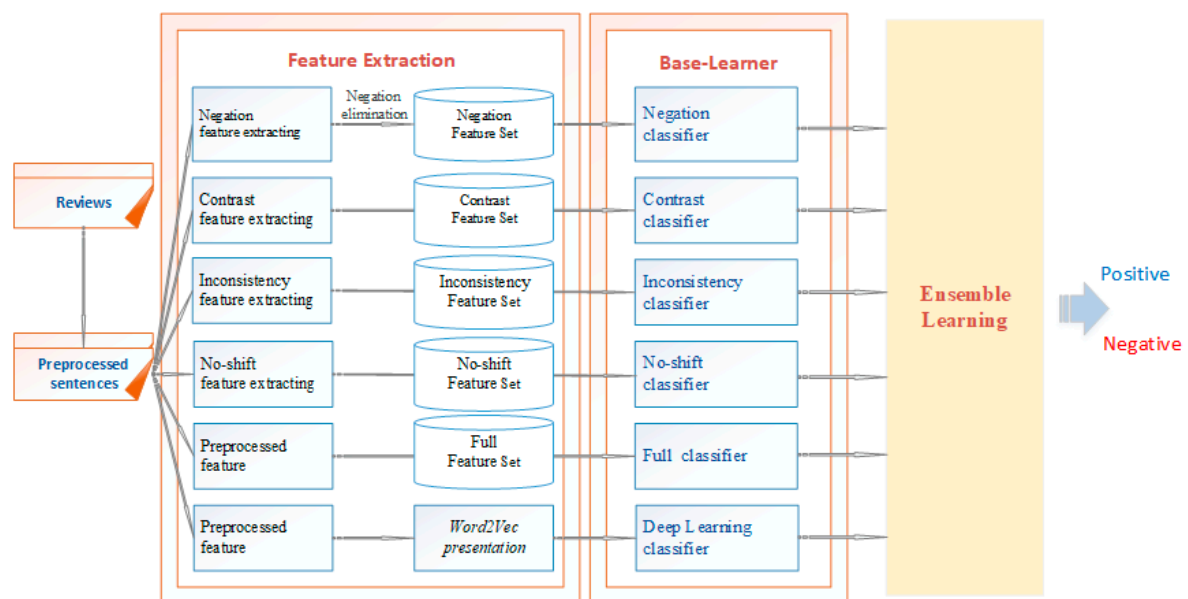
Figure 1. Architecture of sentiment classification system based on ensemble learning model.

3.2. Building Training Datasets for Base Learners

A learning feature representation used to build datasets for base learners is the key task in applying ensemble learning. In this paper, extraction of the following feature types was considered:

- Features of 'surface feature' type: concerning polarity shifting, similar to Rui Xia et al.'s approach [5] (having proposed a feature extracting technique based on rules and statistic method focusing on discovering polarity shifting cases), we define extracting rules according to language characteristics. Text's sentences and clauses will be determined for polarity shifting by negation, contrast,

and inconsistency identifying techniques. The results will be introduced into a corresponding training dataset.

- Features of 'deep feature' type: data which has a complex structure. Therefore, we need to develop automatic extracting methods effectively in order for the system to estimate as well as possible. In traditional machine learning, the feature extracting task is designed and standardized by man, which is a weakness of the previous machine learning methods. Deep learning takes advantage of an architecture of processing multilayers for data component learning representation, as each layer represents different abstract levels of data. We choose deep learning model to extract 'deep features' type.

3.2.1. Extracting 'Surface Feature'

Weighted log-likelihood ratio statistics for sentiment words classification: There are numerous different methods of classifying sentiment words. We apply the weighted log-likelihood ratio statistic method (WLLR) proposed in Rui Xia et al. [5], the WLLR measurement shows a word t_i's correspondence to class c_j through Formula (1):

$$r(t_i, c_j) \; = \; p(t_i, c_j) \log \frac{p(t_i, c_j)}{p(t_i, \bar{c}_j)} \tag{1}$$

where:

$p(t_i, c_j)$: word $t_i's$ *probability* of class c_j and

$p(t_i, \bar{c}_j)$: word $t_i's$ *probability* of another class different from c_j.

- If $r(t_i) > 0$, this word is introduced into the positive sentiment word set, labeled with the measurement $r(t_i)$, and ranked according to its measurement.
- Otherwise, this word is introduced into the negative sentiment word set, labeled with the measurement $|r(t_i)|$, and ranked according to its measurement.

Depending on the word ranking order, we build pairs of sentiment polarity. These pairs will be applied in negation elimination process.

WLLR statistics will be also used to identify sentiment contrast sentence, as Formula (2) indicates:

$$h(s_i) = y \sum_{j=0}^{|s_i|} r(t_j) \tag{2}$$

if $h(s_i) < 0$: inconsistency sentence,
otherwise: no_shift sentence.
with:

- y: label (pos/neg) of the text
- s_i: sentence s_i of the text
- $|s_i|$: number of words in s_i
- $r(t_i)$: given by Formula (1)

Features forming negation dataset: Negation structure is the most popular structure in polarity shifting. Table 3 demonstrates this structure's constant occurrences in the dataset, such as the word "không not" taking place 9778 times in 3,829,253 words in total of the dataset of Vietnamese reviews collected about hotels. Identifying negation structure is realized by checking the occurrences of words in sentences, such as "không not", "chẳng no", "chả don't". These identified sentences will be put into the D_negation set, comprising of negation sentences. After identifying the negation word's positions in the D_negation set's sentences, this negation word will be removed. The first sentiment word having

followed the negation word removed will be replaced with another word bearing contrast sentiment according to Formula (2). Sentiment words which gradually follow will be replaced if they manifest the same sentiment as the first one.

Example: "I do not like this hotel!" will be replaced with "I dislike this hotel!"

Table 3. Statistics realized depending on some negation words' occurrences in Vietnamese language corpus of hotel reviews.

Shifters	Occurrences in the Corpus	
không not		No. of Hits = 9778 File Length (in chars) = 3829253
chẳng no		No. of Hits = 260 File Length (in chars) = 3829253
chả don't		No. of Hits = 7 File Length (in chars) = 3829253

Features forming contrast dataset: Contrast structure is also a popular one in polarity shifting. Table 4 demonstrates the quite important frequency of the word "nhưng but", which occupies 3728 places in 3,829,253 words in total of the dataset of reviews collected about hotels. These words are divided into two groups: the first one is called fore-contrast includes "nhưng but" and "tuy however", and the second one is called post-contrast "mặc_dù although" and "dù though". If a fore-contrast occurs in a sentence, the polarity shifting takes place in the phrase preceding the fore-contrast, and, in the case of post-contrast, sentences which contain post-contrast are shifted themselves. The contrast sentences will be put in the set $D_{contrast}$.

Example: "Khách sạn rất đẹp, vị trí thuận lợi tuy nhiên giá hơi đắt." (The hotel is very nice, its location is good but the price is quite expensive). The polarity shifting occurs in the phrase "Khách sạn rất đẹp, vị trí thuận lợi" (The hotel is very nice, its location is good).

Table 4. Statistics based on some contrast structure words occurring in Vietnamese language corpus of hotel reviews.

Shifters	Occurrences in the Corpus	
mặc_dù although		No. of Hits = 193 File Length (in chars) = 3829253
tuy however		No. of Hits = 1450 File Length (in chars) = 3829253
nhưng but		No. of Hits = 3728 File Length (in chars) = 3829253

Features forming inconsistency dataset: Sentiment inconsistency sentences are the ones which do not demonstrate grammatical polarity shifting but implicate the contrast to sentiment shown in the whole text. This inconsistency is caused by human language, such as implicit, ironical, and satirical sentences. Inconsistency sentences can be identified with the WLLR through evaluating every word in the text. Then, a sentence will be evaluated on polarity shifting with Formula (2).

Relying on the evaluated value, one of the two following decisions will be made:

- If $h(s_i) < 0$, the sentence will be put into the set $D_{inconsistency}$ containing inconsistency data;
- If $h(s_i) \geq 0$, the sentence will be put into the set D_{no_shift} containing unshifted data.

Features in full dataset: Besides, we also use the entire dataset, which is already preprocessed, and use the name D_{full} for another base learner.

As with the strategy of polarity shifting classifying presented above, we have two combined methods for identifying polarity shifting, as shown below: (1) Identifying polarity shifting by rule-based method—building rules and dataset containing words, phrases causing polarity shifting and representative polarity shifting structures in order to identify and remove polarity shifting in sentences; (2) Identifying polarity shifting by statistic method—using WLLR to predict possibilities to shift polarity in sentences. This technique was appropriate for identifying inconsistent sentences, as well as researching and building training models applicable to different domains and languages.

Identifying 'surface features' is described in Figure 2. First, the data was preprocessed. In the next stage, handcrafted features were used that were created using tokenizer and Part-Of-Speech taggers, as well as valance shifter indicators as inputs for the machine learning algorithm. The goal was to produce the most accurate results with 'surface features' that have the highest possibility of prediction.
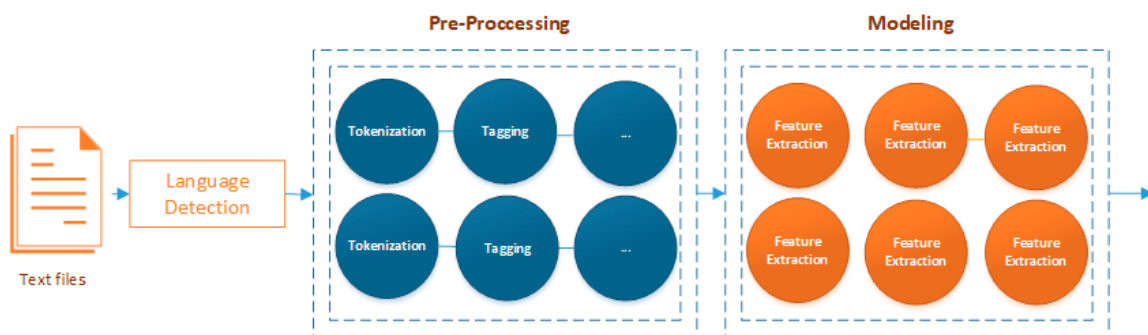

Figure 2. Process identifying 'surface features'.

3.2.2. Extracting Features of 'Deep Feature' Type

Deep learning is a subset of machine learning that depends on learning different multilayers representing of data, each of which automatically transforms the representation at one level into a representation at a higher and more abstract level. The learned representations can be naturally used as features; we call these 'deep features'. Many deep learning models of natural language processing have used input features of word embedding (word vector) [50]—a technique of dense information word learning in a vector space of dense dimensions. Every word is regarded as a point in this space and represented by a vector of constant length. These vectors can represent a language's rules and characteristics. Of word embedding learning models on raw text, Word2Vec is a particularly effective model and usually called for. We take the training dataset of Word2Vec type for inputs into the Long Short-Term Memory (LSTM) network. The LSTM model was introduced by Hochreiter and Schimidhuber [51] and then improved by Gers et al. [52]. LSTM has a similar structure to a Recurrent Neural Network (RNN) [53]. However, instead of only being a one-layer neural network, a state in an LSTM has four layers. The main idea of LSTM is that in each layer there will be a forget gate to decide whether to allow the previously learned information to be used for the current layer or not. Numerous deep learning models that extend the LSTM have been proposed. but the classic LSTM still remains a strong baseline [54].

The highest probability samples belong to positive/negative stages, which were chosen as features of the meta-learner within the ensemble learning classifier, referred to as 'deep features.' The 'deep feature' identification process is described in Figure 3. First, the text data were preprocessed. Then the data were converted into dense vectors using embedded techniques, such as Word2Vec. Next,

the dense vectors were loaded into a deep learning model. The goal was to produce the most accurate results with 'deep features' that have the highest possibility of prediction.
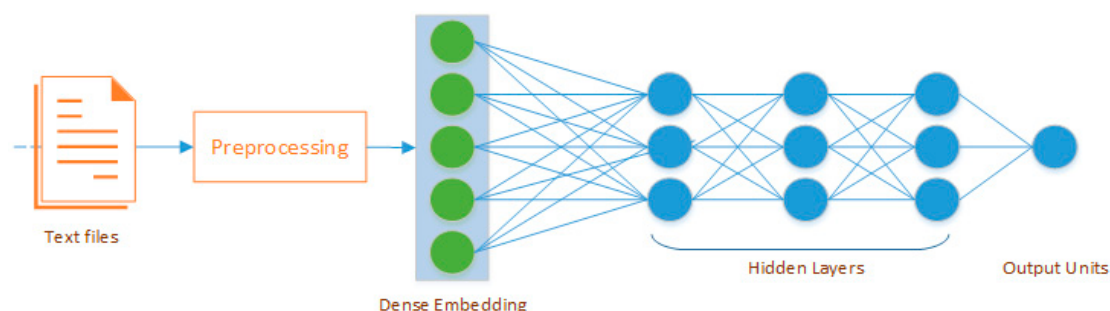

Figure 3. Process identifying 'deep features'.

3.3. Base learners and Meta-Learner

The classic machine learning techniques, like Logistic Regression [55] and Support Vector Machines [56], are used to train datasets that are comprised of features of 'surface feature', which are $D_{negation}$, $D_{contrast}$, $D_{inconsistency}$, D_{no_shift}, and D_{full}. They are the highly estimated techniques for text classifying, in general, and sentiment classifying, in particular. Along with them, a deep learning model is chosen to train all the datasets (D_{full}) for the purpose of identifying features of 'deep feature' for ensemble learning.

Base learners' output values are each sample's probability of belonging to negative and positive stages. These probabilities are used as intensifying learning data for the ensemble stage. Regarding ensemble learning, here are the two models used to blend base-classifiers' results [57]: (1) Rule Fixed model uses fixed rules to choose inputs for ensemble learning, and the majority of ensemble learning's results are based on classifier output's results; (2) Meta-Classifier model found in classifiers' results is taken for ensemble learning model's features. In this paper, we used the Meta-Classifier model with Logistic Regression technique. Figure 4 describes the architecture of ensemble learning using the Meta-Classifier model with base learners' output results, which are each sample's probability of belonging to positive and negative stages.
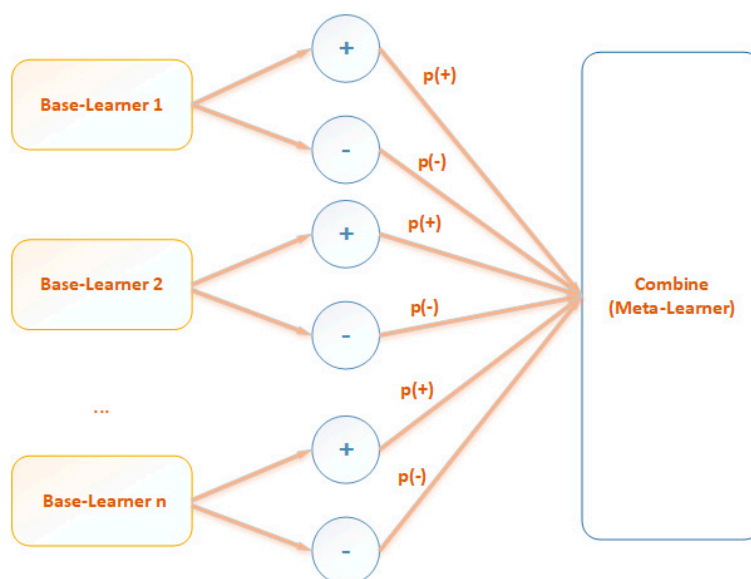

Figure 4. Ensemble learning using Meta-Classifier model.

The whole process can be described in Algorithm 1.

Algorithm 1. Algorithm of classifier ensemble model (CEM)

Input:

- *Dataset $D_{full} = \{d_1, d_2, \ldots, d_n\}$ with:*

  ○ *associated labels set $Y = \{y_1, y_2, \ldots, y_n\}$;*
  ○ *document $d_k = \{s_1, s_2, \ldots s_m\}$ # s_i: a sentence i^{th}; k: 1, $\ldots$,n;*
  ○ *sentence $s_i = \{w_1, w_2, \ldots w_{|si|}\}$ # w_j: a word j^{th}; i=1,..m;*

- *negation indicator $N = \{n_1, n_2, \ldots n_s\}$;*
- *contrast indicator $C = \{c_1, c_2, \ldots c_t\}$;*
- *a base learner algorithm: L;*
- *a deep learning model: DL;*

Process:

Step 1. Identifying these sub datasets: $D_{negation}$, $D_{contrast}$, $D_{inconsistency}$, $D_{no-shift}$:

for k = 1, $\ldots$,n:

for i = 1, $\ldots$ m:

for j = 1, $\ldots$ $|s_i|$:

if $w_{ij} \in N$: put s_i into $d_{k\text{-}negation}$; # capture negations

continue;

if $w_{ij} \in C_1$: put s_{i-1} into $d_{k\text{-}contrast}$; # capture fore-contrast (C_1)

continue;

if $w_{ij} \in C_2$: put s_i into $d_{k\text{-}contrast}$; # capture post-contrast (C_2)

continue;

compute $r(w_{ij})$;   # r is calculated according to the formula (1)

compute $h(s_i)$;   # h is calculated according to the formula (2)

if $h(s_i) < 0$: put s_i into $d_{k\text{-}inconsistency}$

let $d_{k\_no-shift} = d_k - d_{k\_negation} - d_{k\_contrast} - d_{k\_inconsistency}$

let $D_{negation} = \{d_{1\_negation}, d_{2\_negation}, \ldots, d_{n\_negation}\}$;

let $D_{contrast} = \{d_{1\_contrast}, d_{2\_contrast}, \ldots, d_{n\_contrast}\}$;

let $D_{inconsistency} = \{d_{1\_inconsistency}, d_{2\_inconsistency}, \ldots, d_{n\_inconsistency}\}$;

let $D_{no-shift} = \{d_{1\_no-shift}, d_{2\_no-shift}, \ldots, d_{n\_no-shift}\}$;

Step 2. Conducting training phases:

$bl_1 = L(D_{negation})$ # train a base-learner bl_1 on dataset $D_{negation}$

$bl_2 = L(D_{contrast})$ # train a base-learner bl_2 on dataset $D_{contrast}$

$bl_3 = L(D_{no\_shift})$ # train a base-learner bl_3 on dataset D_{no_shift}

$bl_4 = L(D_{inconsistency})$ # train a base-learner bl_4 on dataset $D_{inconsistency}$

$bl_5 = L(D_{full})$ # train a base-learner bl_5 on dataset D_{full}

$bl_6 = DL(W2V(D_{full}))$ # train a deep learning bl_6 on dataset D_{full} with Word2Vec presentation.

Output:

$CEM(d) = argmax_y \in Y \sum_{t=1}^{6} 1 \, (y=bl_t(d))$ # the value of 1(α) is 1 if α is true and 0 otherwise.

---

4. Experiment and Evaluation

4.1. Vietnamese Language

Dataset: We experimented on two datasets containing students' reviews about the university (UIT-VSFC) [58] and reviews about hotels in Vietnam (HOTEL-Reviews). The hotel reviews were posted by users on mytour.vn from 02/8/2010 to 29/6/2017. Review data were preprocessed to remove abbreviations, social network language, signs, logos, etc. Information details about the two datasets are described in Table 5. We proportioned their train-test as 50–50%.

Table 5. Information details about two datasets used in experiment.

Dataset	Trained Data		Experimented Data	
	Number of Positive Reviews	Number of Negative Reviews	Number of Positive Reviews	Number of Negative Reviews
HOTEL-Reviews	932	932	932	932
UIT-VSFC	1285	1285	1285	1285

Models contributing to experiment process:

We compared our model with other strategies, the traditional SVM classification method, along with the deep learning model LSTM, as follows:

- SVM: sentiment classification using the classic machine learning method, Support Vector Machines, with the bag-of-words model, unigram feature.
- LSTM: sentiment classification using Long Short-Term Memory 2×64 hidden-layer units with feature representing the Word2Vec type. The original dimension of our one-hot vector is 74,268, reduced to 300 after performing word embedding. Dropout and recurrent dropout are 0.5. Activation function is sigmoid.
- Classifier ensemble model (CEM)(4C-LR): model comprised of meta-learner using Logistic Regression and four base learners which are contrast learner, inconsistency learner, negation learner, and no_shift learner.
- CEM(5C-LR): model comprised of meta-learner using Logistic Regression and five base learners which are contrast learner, inconsistency learner, negation learner, no_shift learner, and full learner.
- CEM(6C-LR)—the proposed model: model comprised of meta-learner using Logistic Regression and six base learners, which are contrast learner, inconsistency learner, negation learner, no-shift learner, full learner, and LSTM learner.

With "contrast", "inconsistency", "negation", "no-shift", and "full" are the names of sub-datasets, which are described in Section 3.2.1, the Logistic Regression technique is adopted for all base learners.

Experiment results:

Experiment results calculated according to accuracy ratio are shown in Table 6 and Figures 5 and 6.

Table 6. Results experimented on two corpuses of reviews about hotels in Vietnam (HOTEL-Reviews) and UIT-VSFC.

Model	HOTEL-Reviews	UIT-VSFC
SVM	94.15%	92.11%
LSTM	92.75%	94.50%
CEM(4C-LR)	91.84%	90.60%
CEM(5C-LR)	93.07%	96.65%
CEM(6C-LR)	94.36%	97.03%
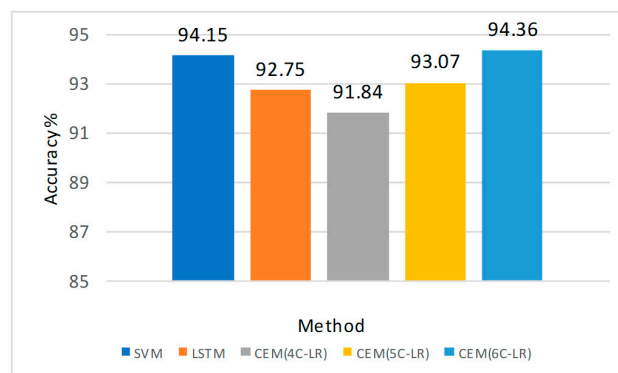
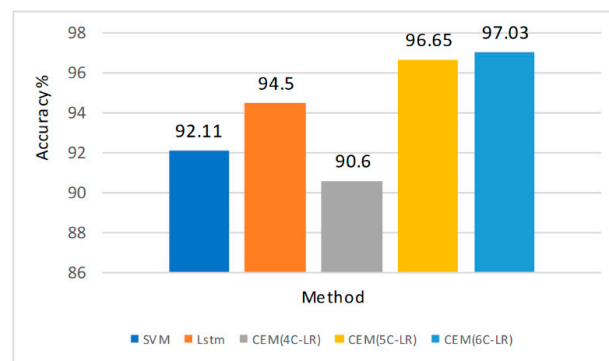**Figure 5.** Results experimented on corpus of HOTEL-Review.

Figure 6. Results experimented on corpus of UIT-VSFC.

4.2. English Language

Dataset: We experimented on corpus proposed by Blitzer et al. [59] including four domains of Electronics, DVD, Books, and Kitchen, each of which contained 1000 reviews labeled positive and 1000 reviews negative. These datasets were used for fair comparison purposes since two other approaches were used, as shown below. We proportioned their train-test as 90–10%.

Models contributing to experiment process:

Similar to the Vietnamese language, the following models contributed to experiment process:

- SVM: sentiment classification using the classic machine learning method Support Vector Machines with the bag-of-words model, unigram feature.
- MLP (Multilayer Perceptron): sentiment classification using neural network with 160 inputs, 2×50 hidden-layer neurons, and 2 outputs. Activation function is softmax.
- PSDEE: method proposed by Rui Xia et al. [5] using four sub-datasets of contrast, inconsistency, negation, and no_shift with unigram feature. There are four base learners for training and combining tasks.
- LSS: method proposed by Shoushan et al. [43] using two sub-datasets of shift and no_shift with unigram feature. There are two base learners for training and combining tasks.
- CEM(5C-LR): model comprising of meta-learner using Logistic Regression and five base learners which are contrast learner, inconsistency learner, negation learner, no_shift learner, and full learner.
- CEM(6C-LR)—the proposed model: model comprising of meta-learner using Logistic Regression and six base learners which are contrast learner, inconsistency learner, negation learner, no_shift learner, full learner, and MLP learner.

Experiment results:

Experiment results, which were calculated according to accuracy ratio are shown in Table 7 and Figures 7–10.

Table 7. Results experimented on four corpuses of Electronics, DVD, Books, and Kitchen.

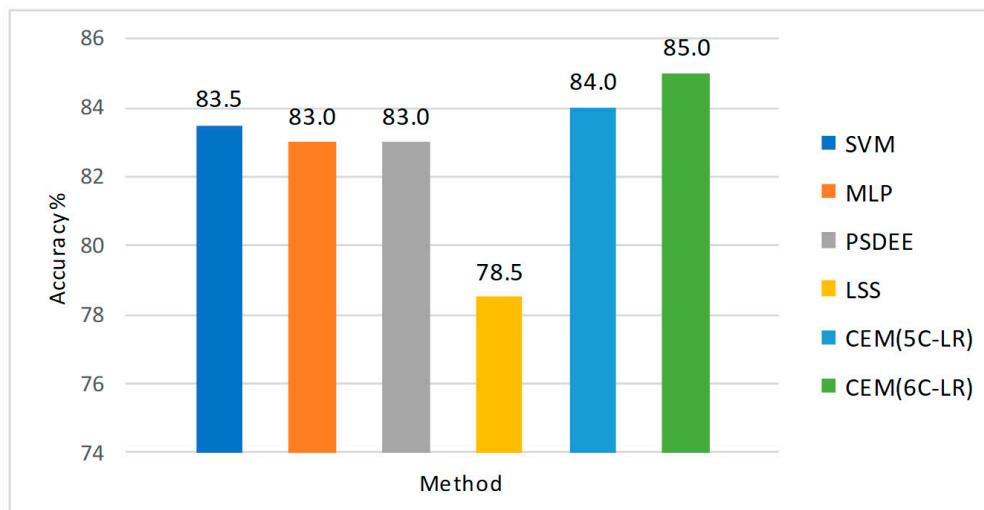| Model	Electronics	DVD	Books	Kitchen
SVM	83.50%	85.00%	80.00%	84.50%
MLP	83.00%	82.00%	73.00%	86.50%
PSDEE [6]	83.00%	81.00%	80.04%	85.70%
LSS [34]	78.50%	77.00%	83.00%	84.90%
CEM(5C-LR)	84.00%	83.50%	81.00%	84.50%
CEM(6C-LR)	85.00%	85.50%	81.00%	87.00%

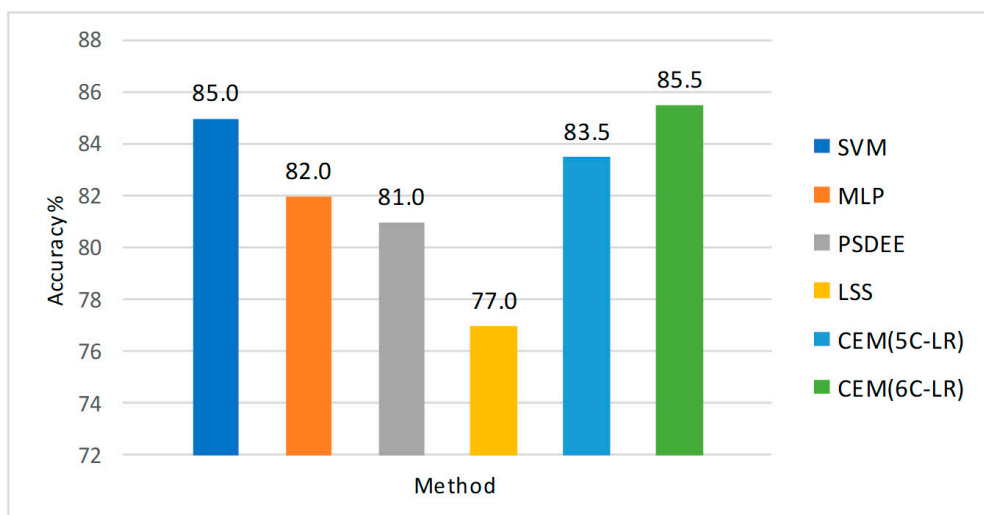**Figure 7.** Results experimented on corpus Electronics.

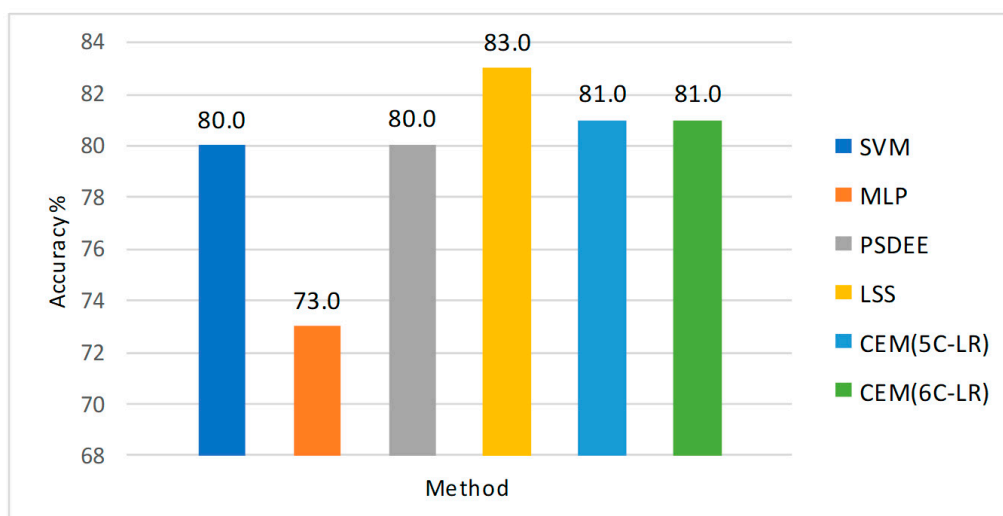**Figure 8.** Results experimented on corpus DVD.

Figure 9. Results experimented on corpus Books.
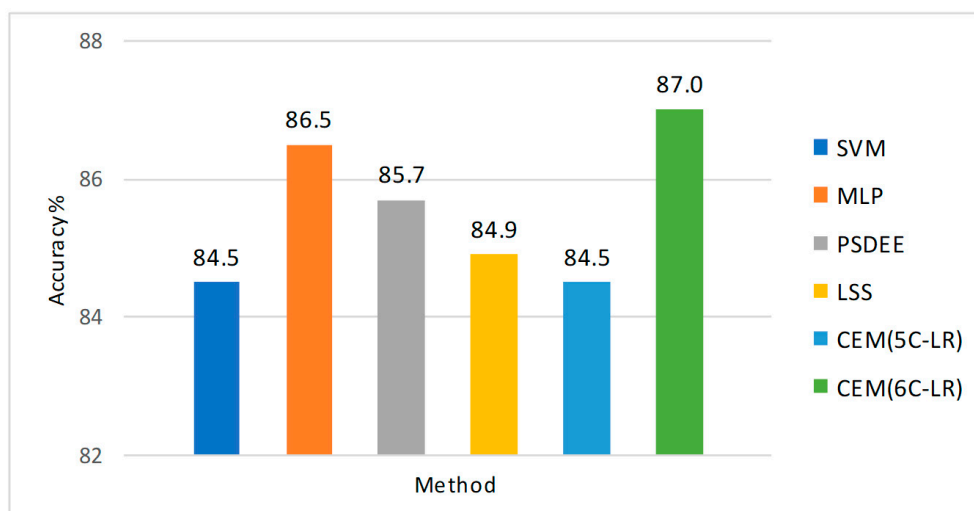
Figure 10. Results experimented on corpus Kitchen.

4.3. Evaluation

Based on the experiment results for the Vietnamese language, we have come to the following conclusions:

- Models interested in 'deep features', such as the CEM(6C-LR), leads to results better than other models in both datasets experimented, especially compared with other strategies of four base learners, five base learners, and the baseline machine learning method SVM.
- Ensemble learning of training sets containing 'deep feature' and features relevant to polarity shifting of 'surface feature' type leads to classification results better than the deep learning state-of-the-art model in sentiment classification (LSTM). The 'deep features' were the most effective features used to classify sentiment in an ensemble system.
- Dataset size also has an effect on every method's effectiveness. With limited data (HOTEL-Reviews), SVM always proves to be an effective text classification method when compared with LSTM or the model proposed by us (CEM(6C-LR)).

Our dataset size in the English language is limited, and, therefore, we have chosen the neural multilayer network (MLP) model instead of LSTM. It is certain that the original LSTM suffers from poor performance when applied with a small dataset. Experiment results show that the model proposed by us, CEM(6C-LR), attains higher classification effectiveness than other methods, especially when compared with the PSDEE method given by Xia et al. [5] and the LSS method proposed by Shoushan et al. [43]. Figures 11 and 12 show the average accuracy of each base learner in the method of Xia et al. [5] and in the proposed model.

The Xia et al. method used a statistical approach to predict data test labels in the inconsistency learner and no-shift learner. The average accuracy of the two learners was only 70%, with a final accuracy of 82.43%. The proposed model used the baseline method for the inconsistency learner and no-shift learner, which provided an average accuracy of 76%. By adding a deep learning learner, the proposed system achieved a final accuracy of 84.62%. The learners that contributed the most to the accuracy of the system were the full learner, the no-shift learner, and the deep learning learner, corresponding to no-shift features that did not cause sentiment shifts, and 'deep features'.

The WLLR statistical model was not able to perform well with the Vietnamese language. The WLLR's classification of polarized words was insufficient, and data alone cannot increase the accuracy of polarized words. This approach revealed its drawbacks when dealing with grammar that had a complex semantic structure, such as Vietnamese. For instance, a well-known example error was classifying the word pair "không_thích $_{don't\ like}$" and "ghét $_{hate}$" into the same ranking. However, in the

English language, "don't like" may replace "dislike." This can be solved by replacing the WLLR method with the emotional dictionary proposed by Tran et al. [31] to rank pairs of words bearing sentiment.

The reasons for our proposed method achieving acceptable results can be summarized as follows:

- Our model can capture various cases of sentiment shifts and introduce appropriate treatment for each. In improving on Rui Xia et al.'s approach, we have built additional base learners for various sub-datasets that improved system performance.
- Our model uses deep learning to automatically learn features that are implicit as input to the meta-learner.
- Our model proved the powerful ensemble learning suite by having datasets with features of different characteristics.
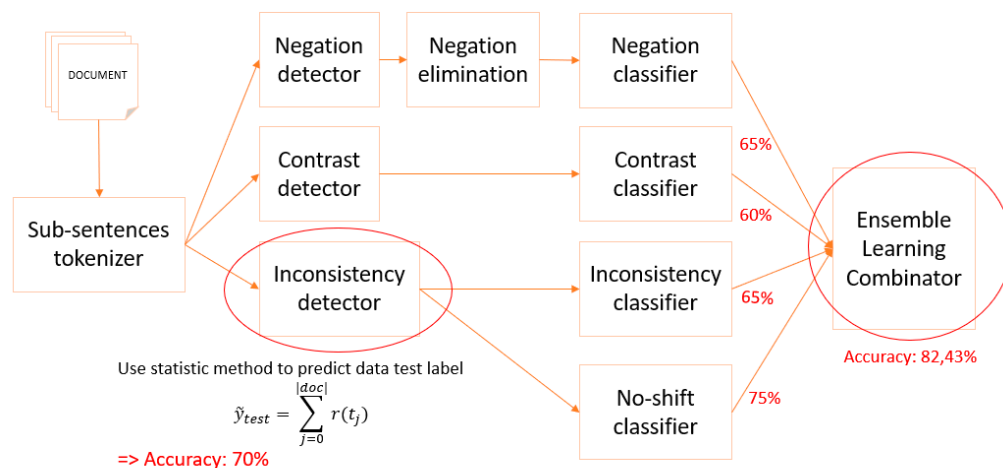

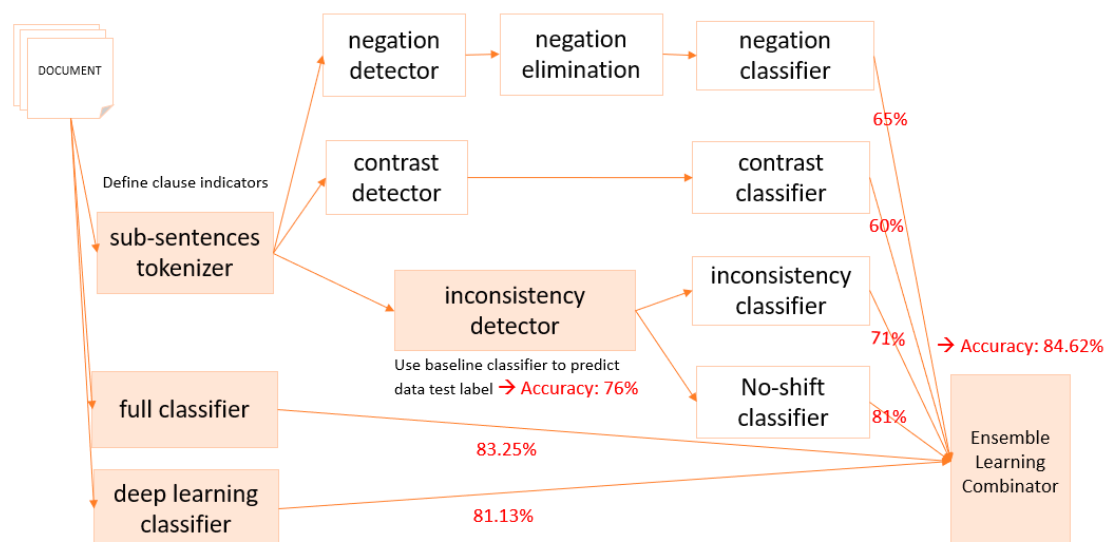**Figure 11.** Average accuracy of the Xia et al. approach.

Figure 12. Average accuracy of CEM(6C-RL).

5. Conclusions and Future Research Plan

In this paper, we introduced a novel model that integrates the advantages of deep learning, machine learning, statistics, and rule-based techniques. Although the computational cost of the proposed system is higher than the compared algorithms, the system bears multiple distinctive characteristics. We combined different methods, identified polarity shifting based on language structures and techniques, and used a word-embedding model with deep learning. This approach

captures both 'surface features' and 'deep features' in text and allows our system to achieve results better than other models. In addition, experiments have demonstrated that the proposed model works effectively with other languages, such as English.

Our future work will focus on analyzing and conducting more experiments on deep learning as well as polarity-shifting structures in texts with the objective of uncovering the limits (if that is the case) of existing models when applied to different datasets and domains in response to the sentiment classification of reviews on social networks (especially Twitter and Facebook), which presently capture the attention of the maximum population. Texts of this kind, which are often short and written using complex structures, alter the meanings of sentences and render their identification more difficult. One solution to this problem could involve broadening the rules and the vocabulary corpus to process these specific sentences. The intention to experiment with other deep learning models to strengthen the system and enhance its accuracy will be taken account of. As mentioned in the evaluation, replacing the WLLR statistical method with the VNSD dictionary [31] was considered for sentiment score ranking.

References

1. Li, G.; Liu, F. Sentiment analysis based on clustering: A framework in improving accuracy and recognizing neutral opinions. *Appl. Intell.* **2014**, *40*, 441–452. [CrossRef]
2. Dave, K.; Lawrence, S.; Pennock, M.D. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In Proceedings of the 12th International Conference on World Wide Web, Budapest, Hungary, 20–24 May 2003; p. 519.
3. Nasukawa, T.; Yi, J. Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the 2nd International Conference on Knowledge Capture, Sanibel Island, FL, USA, 23–25 October 2003; p. 70.
4. Tang, D.; Qin, B.; Liu, T. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 19–21 September 2015; pp. 1422–1432.
5. Xia, R.; Xu, F.; Yu, J.; Qi, Y.; Cambria, E. Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. *Inf. Process. Manag.* **2016**, *52*, 36–45. [CrossRef]
6. Marcheggiani, D.; Täckström, O.; Esuli, A.; Sebastiani, F. Hierarchical Multi-Label Conditional Random Fields for Aspect-Oriented Opinion Mining. In *Lecture Notes in Computer Science (Including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Cham, Switzerland, 2014; Volume 8416 LNCS, pp. 273–285.
7. Yang, B.; Cardie, C. Context-aware Learning for Sentence-level Sentiment Analysis with Posterior Regularization. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27 June 2014; Volume 1, pp. 325–335.
8. Chinsha, T.C.; Joseph, S. A syntactic approach for aspect based opinion mining. In Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing, Anaheim, CA, USA, 7–9 February 2015; pp. 24–31.
9. Tran, T.K.; Phan, T.T. Mining opinion targets and opinion words from online reviews. *Int. J. Inf. Technol.* **2017**, *9*, 239–249. [CrossRef]
10. Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, 6–7 July 2002; Volume 10, pp. 79–86.
11. Riaz, S.; Fatima, M.; Kamran, M.; Nisar, M.W. Opinion mining on large scale data using sentiment analysis and k-means clustering. *Cluster. Comput.* **2017**, 1–16. [CrossRef]

12. Wang, G.; Zheng, D.; Yang, S.; Ma, J. FCE-SVM: A new cluster based ensemble method for opinion mining from social media. *Inf. Syst. e-Bus. Manag.* **2017**, *16*, 1–22. [CrossRef]

13. Turney, P.D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 417–424.

14. Muhammad, A.; Wiratunga, N.; Lothian, R. Contextual sentiment analysis for social media genres. *Knowl.-Based Syst.* **2016**, *108*, 92–101. [CrossRef]

15. Khan, F.H.; Qamar, U.; Bashir, S. Lexicon based semantic detection of sentiments using expected likelihood estimate smoothed odds ratio. *Artif. Intell. Rev.* **2017**, *48*, 113–138. [CrossRef]

16. Balahur, A.; Hermida, J.M.; Montoyo, A. Detecting implicit expressions of emotion in text: A comparative analysis. *Decis. Support Syst.* **2012**, *53*, 742–753. [CrossRef]

17. Keshavarz, H.; Abadeh, M.S. ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs. *Knowl.-Based Syst.* **2017**, *122*, 1–16. [CrossRef]

18. Severyn, A.; Moschitti, A.; Uryupina, O.; Plank, B.; Filippova, K. Multi-lingual opinion mining on YouTube. *Inf. Process. Manag.* **2016**, *52*, 46–60. [CrossRef]

19. Hajmohammadi, M.S.; Ibrahim, R.; Selamat, A. *Graph-Based Semi-supervised Learning for Cross-Lingual Sentiment Classification*; Springer: Cham, Switzerland, 2015; pp. 97–106.

20. Claypo, N.; Jaiyen, S. Opinion mining for thai restaurant reviews using K-Means clustering and MRF feature selection. In Proceedings of the 7th International Conference on Knowledge and Smart Technology (KST), Chonburi, Thailand, 28–31 January 2015; pp. 105–108.

21. Saif, H.; He, Y.; Fernandez, M.; Alani, H. Contextual semantics for sentiment analysis of Twitter. *Inf. Process. Manag.* **2016**, *52*, 5–19. [CrossRef]

22. Vulić, I.; De Smet, W.; Tang, J.; Moens, M.F. Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Inf. Process. Manag.* **2015**, *51*, 111–147. [CrossRef]

23. Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-Based Methods for Sentiment Analysis. *Comput. Linguist.* **2011**, *37*, 267–307. [CrossRef]

24. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]

25. Wu, H.; Gu, Y.; Sun, S.; Gu, X. Aspect-based Opinion Summarization with Convolutional Neural Networks. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 24–29 July 2016; pp. 3157–3163.

26. Jianqiang, Z.; Xiaolin, G.; Xuejun, Z. Deep Convolution Neural Networks for Twitter Sentiment Analysis. *IEEE Access* **2018**, *6*, 23253–23260. [CrossRef]

27. Polanyi, L.; Zaenen, A. *Contextual Valence Shifters. Computing Attitude and Affect in Text: Theory and Applications*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 1–10.

28. Tran, T.K.; Phan, T.T. *Computing Sentiment Scores of Adjective Phrases for Vietnamese*; Springer: Cham, Switzerland, 2016; pp. 288–296.

29. Tran, T.K.; Phan, T.T. Computing Sentiment Scores of Verb Phrases for Vietnamese. In Proceedings of the 28th Conference on Computational Linguistics and Speech Processing (ROCLING 2016), Tainan, Taiwan, 10 November 2016; pp. 204–213.

30. Tran, T.K.; Phan, T.T. Toward Contextual Valence Shifters in Vietnamese Reviews. In Proceedings of the 29th Conference on Computational Linguistics and Speech Processing (ROCLING 2017), Taipei, Taiwan, 27–28 November 2017; pp. 152–159.

31. Tran, T.K.; Phan, T.T. A hybrid approach for building a Vietnamese sentiment dictionary. *J. Intell. Fuzzy Syst.* **2018**, *35*, 967–978. [CrossRef]

32. Araque, O.; Corcuera-Platas, I.; Sánchez-Rada, J.F.; Iglesias, C.A. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Syst. Appl.* **2017**, *77*, 236–246. [CrossRef]

33. Xu, G.; Huang, C.R. Extracting Chinese polarity shifting patterns from massive text corpora. *Ling. Sin.* **2016**, *2*, 5. [CrossRef]

34. De Albornoz, J.C.; Plaza, L.; Gervás, P. A hybrid approach to emotional sentence polarity and intensity classification. In Proceedings of the Fourteenth Conference on Computational Natural Language Learning, Uppsala, Sweden, 15–16 July 2010; pp. 153–161.

35. Jia, L.; Yu, C.; Meng, W. The effect of negation on sentiment analysis and retrieval effectiveness. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, China, 2–6 November 2009; pp. 1827–1830.

36. Domingos, P.M. A few useful things to know about machine learning. *Commun. ACM* **2012**, *55*, 78. [CrossRef]

37. Verma, A.; Mehta, S. A comparative study of ensemble learning methods for classification in bioinformatics. In Proceedings of the 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence, Noida, India, 12–13 January 2017; pp. 155–158.

38. Xie, G.; Zhao, Y.; Jiang, M.; Zhang, N. A Novel Ensemble Learning Approach for Corporate Financial Distress Forecasting in Fashion and Textiles Supply Chains. *Math. Probl. Eng.* **2013**, *2013*, 493931. [CrossRef]

39. Li, Y.; Bai, C.; Reddy, C.K. A Distributed Ensemble Approach for Mining Healthcare Data under Privacy Constraints. *Inf. Sci.* **2016**, *330*, 245–259. [CrossRef] [PubMed]

40. Xia, R.; Zong, C.; Li, S. Ensemble of feature sets and classification algorithms for sentiment classification. *Inf. Sci.* **2011**, *181*, 1138–1152. [CrossRef]

41. Wen, L.; Weili, W.; Yuefeng, C. Heterogeneous Ensemble Learning for Chinese Sentiment Classification. *J. Inf. Comput. Sci.* **2012**, *9*, 4551–4558.

42. Su, Y.; Zhang, Y.; Ji, D.; Wang, Y.; Wu, H. *Ensemble Learning for Sentiment Classification*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 84–93.

43. Li, S.; Lee, S.Y.M.; Chen, Y.; Huang, C.R.; Zhou, G. Sentiment classification and polarity shifting. In Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, 23–27 August 2010; pp. 635–643.

44. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2015; Volume 521, p. 436.

45. Johnson, R.; Zhang, T. Effective Use of Word Order for Text Categorization with Convolutional Neural Networks. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, CO, USA, 31 May–5 June 2015; pp. 103–112.

46. Li, Q.; Jin, Z.; Wang, C.; Zeng, D.D. Mining opinion summarizations using convolutional neural networks in Chinese microblogging systems. *Knowl.-Based Syst.* **2016**, *107*, 289–300. [CrossRef]

47. Zhou, X.; Wan, X.; Xiao, J. Attention-based LSTM Network for Cross-Lingual Sentiment Classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 247–256.

48. Nguyen, D.; Vo, K.; Pham, D.; Nguyen, M.; Quan, T. *A Deep Architecture for Sentiment Analysis of News Articles*; Springer: Cham, Switzerland, 2018; pp. 129–140.

49. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.

50. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.

51. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]

52. Gers, F. Long Short-Term Memory in Recurrent Neural Networks. Ph.D. Dissertation, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2001.

53. Jain, L.C.; Medsker, L.R. *Recurrent Neural Networks: Design and Applications*, 1st ed.; CRC Press Inc.: Boca Raton, FL, USA, 1999.

54. Melis, G.; Dyer, C.; Blunsom, P. On the State of the Art of Evaluation in Neural Language Models. In Proceedings of the 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.

55. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd ed.; Chapman & Hall: London, UK, 1989.

56. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]

57. Freund, Y.; Schapire, R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]

58. Ngan, N.L.T.; Kiet, V.N.; Vu, D.N.; Phu, X.V.N.; Tham, T.H.T. UIT-VSFC: Vietnamese Students' Feedback Corpus for Sentiment Analysis. In Proceedings of the 10th International Conference on Knowledge and Systems Engineering (KSE), Ho Chi Minh City, Vietnam, 24–26 October 2019.

59. Blitzer, J.; Dredze, M.; Pereira, F. *Domain Adaptation for Sentiment Classification*; Association of Computational Linguistics (ACL): Stroudsburg, PA, USA, 2007.