Computers, Materials & Continua
DOI: 10.32604/cmc.2023.033752
Article

Using Informative Score for Instance Selection Strategy in Semi-Supervised Sentiment Classification

Vivian Lee Lay Shan, Gan Keng Hoon*, Tan Tien Ping and Rosni Abdullah

School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang, 11800, Malaysia
*Corresponding Author: Gan Keng Hoon. Email: khgan@usm.my
Received: 27 June 2022; Accepted: 22 September 2022

Abstract: Sentiment classification is a useful tool to classify reviews about sentiments and attitudes towards a product or service. Existing studies heavily rely on sentiment classification methods that require fully annotated inputs. However, there is limited labelled text available, making the acquirement process of the fully annotated input costly and labour-intensive. Lately, semi-supervised methods emerge as they require only partially labelled input but perform comparably to supervised methods. Nevertheless, some works reported that the performance of the semi-supervised model degraded after adding unlabelled instances into training. Literature also shows that not all unlabelled instances are equally useful; thus identifying the informative unlabelled instances is beneficial in training a semi-supervised model. To achieve this, an informative score is proposed and incorporated into semi-supervised sentiment classification. The evaluation is performed on a semi-supervised method without an informative score and with an informative score. By using the informative score in the instance selection strategy to identify informative unlabelled instances, semi-supervised models perform better compared to models that do not incorporate informative scores into their training. Although the performance of semi-supervised models incorporated with an informative score is not able to surpass the supervised models, the results are still found promising as the differences in performance are subtle with a small difference of 2% to 5%, but the number of labelled instances used is greatly reduced from 100% to 40%. The best finding of the proposed instance selection strategy is achieved when incorporating an informative score with a baseline confidence score at a 0.5:0.5 ratio using only 40% labelled data.

Keywords: Document-level sentiment classification; semi-supervised learning; instance selection; informative score

1 Introduction

A customer review is a textual review of a product or service made by a customer who had an experience with the product or service. The purpose served by these reviews is to share one's opinions towards the product or service and feedback one opinion to the seller. Reviews contain a wealth of

information about sentiments and attitudes towards a product or service. These reviews reflect the point of view and experiences of the reviewer, which are possibly helpful for potential buyers making their purchasing decisions and for businesses to better understand how the customers feel about their products. Reference [1] reported that there are 73% to 87% of purchase decisions among the online reviews' readers are greatly affected by the reviews of various services such as restaurants and hotels. However, manual analysis of a great number of opinions is very difficult, time-consuming, and in some cases impossible. Therefore, sentiment analysis has been introduced to discover the knowledge through expressed comments in an effective way that can never be achieved with manual analysis. Some of the pioneer works [1,2] successfully applied sentiment analysis to examine and analyse opinions within the text. Sentiment classification is one of the tasks within the sentiment analysis process; its purpose is to classify the sentiment of a user's opinion towards a target that is expressing positive or negative polarity [3]. There are two groups of methods for performing sentiment classification, lexicon-based methods and machine learning-based methods. Most of these methods require annotated input, which is costly and labor-intensive, for model training. Currently, manual annotation is the most common way of acquiring high-quality annotated input, the work is still manageable if the data size is small (up to 3000). On the other hand, training with deep learning usually requires the data size to be at least 10,000 which is a huge number and not practical to annotate manually.

Semi-supervised methods are possible alternatives for reducing the need for annotated input. They use a combination of a small part of labelled input and a large part of unlabeled input for model training. Compared to supervised machine learning methods, labour force and cost in acquiring labelled input are reduced as only a small number of labelled inputs is needed for training a semi-supervised model. Moreover, the accuracy of semi-supervised models is comparable to a supervised model. Nevertheless, some works reported that the performance of the semi-supervised model degraded after adding unlabeled instances with a predicted label into training [4,5,6]. The suggested solution by some works is to add in only informative unlabeled instances that have positive impacts on model performances [7,8]. Understanding what informative unlabeled instances in the semi-supervised training context help to reduce the time taken to train a good model and avoids wasting time in finding the unlabeled instances that can positively impact the model performance. Nevertheless, there are fewer discussions on what informative unlabeled instances are in semi-supervised training.

Along with the success of semi-supervised learning in many other application domains such as computer vision, semi-supervised methods are gaining enormous attention in the sentiment classification community. The methods receive positive encouragement, but some studies pointed out that although semi-supervised sentiment classification can work with a limited amount of labelled text, however, observed a slow decline in accuracy of the model after augmenting pseudo-labelled into each training iteration ([4,9,10]). A suggested solution from studies is to include only informative unlabeled data that have a positive impact in the training of semi-supervised sentiment classification ([7,8,11]). However, current methods have substantially focused on selecting confidently predicted instances only. On the other hand, proposed methods that work on selecting informative predicted instances require high computational resources [11,12]. This necessitates the need for a simple scoring formula that can represent the informativeness of the unlabeled data.

Semi-supervised sentiment classification methods require only a small portion of a labelled dataset for model training. But the ratio of labelled and unlabeled data is rarely reported or suggested in studies. Without a guide, it is difficult to estimate when to stop collecting labelled text and this makes it difficult for resource management. Therefore, a suggestion on the optimal ratio of labelled and unlabeled data is required to minimize the time and effort in acquiring labelled text for semi-supervised training. In this paper, the research questions that will be addressed are as follows.

1) Can context-related information be used to improve the instance selection process in semi-supervised sentiment classification?
2) What is the optimal ratio of labelled and unlabeled data for semi-supervised sentiment classification?
3) What are the optimal parameters for reviewing the informative score and instance selection strategy proposed?

The main contribution of our proposed method is semi-supervised sentiment classification with a review informative score. The review informative score enables evaluation of unlabelled reviews, checking their informativeness, and brings positive impacts to semi-supervised model performances. Moreover, the proposed instance selection strategy can select the confidently predicted and informative predicted instances. This allows the community to move towards creating powerful semi-supervised or even unsupervised sentiment classification models with satisfactory performances. Our research also suggested an optimal ratio of labelled and unlabelled data in semi-supervised model training, making it possible for semi-supervised model users to roughly estimate the number of labelled texts they should be collecting. Besides, the proposed methodology also automates the data annotation using only a small amount of labelled data and the results are comparably good to supervised models. This in turn allows for reducing dependency on supervised models which require fully annotated inputs that are costly to acquire. The remainder of the paper is structured as follows. Section 2 states the related works of semi-supervised sentiment classification. Section 3 explained the proposed methodology. Section 4 describes the experimental setups and results. Finally, Section 5 draws the conclusion and the findings of the paper.

2 Related Works

Recently, semi-supervised learning techniques have been vastly applied to improve the efficiency of sentiment classification. Semi-supervised sentiment classification studies mainly aim to reduce the burden of acquiring labelled datasets while attaining high classification accuracy. The discussions of related works are presented from two perspectives, i.e., semi-supervised approaches and instance selection strategy.

2.1 Semi-Supervised Approaches

One of the approaches is to utilize unsupervised learning in deriving new knowledge from unlabeled instances or to use publicly available knowledge bases to improve semi-supervised model performance. Reference [13] integrated unsupervised knowledge derived from unlabeled instances into model training. Reference [14] performed semi-supervised sentiment analysis using revised sentiment scores based on SentiWordNet. Whereas [15] extended the number of labelled data using an unsupervised joint sentiment/topic model and filter confidently predicted instances for better model performance.

Another approach is to make use of different views of the data. This approach usually pairs together with co-training, one of the popular semi-supervised methods, which involves two or more classifiers for sentiment classification. Reference [16] described how they use word embedding and character-based embedding of forum posts to improve the accuracy of sentiment classification. The two views of embedding ensure the deep neural network can extract different features from texts. Besides, the proposed double-check strategy is applied to select samples with the same pseudo-labels from both classifiers. Their experimental results demonstrate the effectiveness of their proposed methods for training models on limited labelled data. Note that training multiple classifiers requires

more resources than training only one. In addition, graph-based models and generative models are also commonly used semi-supervised methods for solving challenges in sentiment analysis. Reference [17] employed graph-based text representation for sentiment analysis whereas [18] utilized a generative model to perform sentiment feature extraction for classification.

2.2 Instance Selection Strategy

Semi-supervised methods are undeniably powerful and competent. But due to its model predictions being used in the training process iteratively; thus it is hard to guarantee the introduction of unlabeled data will not degrade performance. In extreme cases, the classifier will classify all the unlabeled data into one of the undesirable classes. Hence, studies involved the inclusion of only informative pseudo-labelled instances that have positive impacts on classifiers (see Table 1)

Table 1: Example product metadata and product review

Research work	Learning method	Instance selection strategy	Target
Semi-supervised sentiment analysis based on dynamic threshold and multi-classifiers [12]	SVM	Dynamic threshold based on quality and quantity of auto-labelled training data	Improve accuracy of semi-supervised sentiment analysis by extending labelled data with dynamic threshold and multi-classifiers
Using sentiment labelling for extending labelled data for semi-supervised sentiment classification [15]	The unsupervised joint sentiment/topic model combines with semi-supervised training	Confidence-based and class-balanced instances	Extend the number of labelled data and improve self-training model performance by filtering confidently predicted instances
Evidence-based uncertainty sampling for active learning [19]	NB	Evidence-based uncertainty	Improve performance of learning by uncovering the reasons for a model's uncertainty
Self-training with selection-by-rejection [11]	Logistic regression and SVM	Rejection strategy	Improves performance of self-training by decreasing the disagreement region of hypotheses

Reference [12] proposed dynamic thresholds for different iterations to maximize the number of accurately labelled data selected. The proposed dynamic threshold is based on two factors, the quality and the quantity of pseudo-labelled training data. Evaluation of predicted label quality is related to specific prediction models. The SVM classifier was adopted in this work; thus the quality of predicted labels is defined as the distance between the pseudo-labelled instance and the hyperplane. For example, pseudo-labelled instances A and B fall on the side of the positive hyperplane. The label quality of sample B is higher than sample A as B is much further away from the hyperplane. They proposed to set a high threshold in former iterations and decrease the threshold when the iteration number

increases. A higher threshold for the first few iterations is proposed as the quantity of initial labelled training data is small. High-quality pseudo-labelled data is preferred to prevent the deterioration of classifier performance at later iterations. At later iterations, the threshold is lower to guarantee enough labelled data.

Reference [15] selected a set of class-balanced and confidently predicted instances to be included in the next iteration of classifier training. The predicted instances are first ranked according to confidence, and then a set of top N instances is selected. The selected set is further divided into two disjoint sets based on the sentiment polarities of each pseudo-labelled instance. The set with more instances is taken as the major class, whereas the other set is taken as the minor class. Instances from both classes are sampled based on the number of instances in the minor class to form a class-balanced set. These class-balanced and confidently predicted instances are included in the next training iteration and subtracted from the unlabeled dataset.

Reference [11] proposed a self-training algorithm that decreases the disagreement region of hypotheses based on three properties of informative unlabeled data. Informative unlabeled data improve classifier performance if they provide additional information on the true decision boundary, retain the overall data distribution and introduce bounded noise. They transform the problem of selecting a set of informative unlabeled data into the problem of rejecting the same set of unlabeled data with its labels inverted. This is due to when the inclusion of an unlabeled dataset has a marginal impact on the classification of labelled data, it is difficult to conclude whether the classifier is approaching the true decision boundary. When the addition of unlabeled data introduced a much greater misclassification of the labelled data, the unlabeled data with currently assigned labels can be confidently recognized as harmful, because they are incorrectly labelled, or they have been sampled disproportionately. Furthermore, not all unlabeled data helps improve the classifier's performance despite being correctly labelled. Thus, only the subset of unlabeled data that deteriorates the performance of a new classifier trained using the training set including the same set of unlabeled data with inverted labels, will be added to the training set.

In [19], the authors proposed an evidence-based framework that can uncover the reasons why a model is uncertain in a given instance. Two reasons for uncertainty in a model are discussed. A model can be uncertain about an instance because it has strong but conflicting evidence for both classes, introduced as conflicting-evidence instances. On the other hand, a model can be uncertain about an instance because it does not have enough evidence for either class or known as insufficient-evidence instances. Their work found that the conflicting-evidence instances significantly improved the learning efficiency of a model, whereas the insufficient-evidence instances provided the least value to a model.

2.3 Gap Analysis

Based on the literature review shows, there are limited studies on instance selection strategies to improve the performances of semi-supervised sentiment classification models. The focus is commonly on unsupervised pre-trained networks and external knowledge bases. Besides, instance selection strategies proposed are predominantly to identify confidently predicted instances instead of informative instances.

Quality and the quantity of pseudo-labelled data are two key factors in the dynamic threshold proposed by [12]. The quality of pseudo-labelled instances is defined accordingly to the prediction model used. However, the high and low thresholds mentioned in the research were not defined or suggested. The instance selection strategy proposed by [15] is simple and easily adapted to other works, but the threshold value suggested is extensively tested on the joint sentiment/topic model only.

Adapting the proposed instance selection strategy requires the examination of an optimal confidence threshold value. These proposed methods identify confidently predicted instances only. Three properties of informative unlabelled data are introduced by [11] to provide additional information on the true decision boundary, retain the overall data distribution, and introduce bounded noise. Nevertheless, the method proposed is not efficient because it involves training and testing all unlabelled subsets, reverting the label back and forth. This requires high computational resources for model training. All this leads to the need for an efficient method to discover informative data points that contribute positively to model performance.

Besides, the ratio of the number of labelled and unlabelled training instances used in experiments is rarely reported in studies, even though semi-supervised learning has been used extensively in sentiment classification. This leads to resources such as time and money being not utilized to their fullest as there is no guidance on the estimated number of labelled texts required in the training. Therefore, a suggestion on the optimal ratio of labelled and unlabelled instances is required.

3 Method

This section presents the proposed research method of semi-supervised classification with the purpose ability to identify informative unlabeled instances. We first introduce sample data, followed by details of the flow of semi-supervised learning.

3.1 Preliminary

An example of contents from the datasets, i.e., product metadata and product review, is shown in Table 2 to guide the explanation of the proposed method in the subsequent section. productID acts as the key attribute that links the relationship between product metadata and product review.

Table 2: Example product metadata and product review

	Attributes	Description	Value
Product metadata	productID	ID of product	"0000069512"
	productTitle	Name of product	"Refrigerator storage organizer"
	productFeature	Features of product	["Great organizer for fridge", "Easy to carry", "Clear material"]
	productDesc	Description of product	"This organizer is great for fridge. Cutout side handles for easy carry. Clear material enable easy see through what are stored inside the bin."
Product review	helpful	Helpful votes of review	3
	review	Text of review	"Love these bins to help me keep my fridge organized. These bins not only has helped me see what I have but makes me happy seeing how tidy it looks now too!"
	overall	Rating of product	5.0
	reviewTime	Time of review	"01 28, 2009"

3.2 The Flow of Semi-Supervised Classification

Given a dataset R consists of n review entries, after pre-processing and matching with their respective product ID, the product metadata (product title, product features, and product description), review posted date and helpful votes (similar to upvotes and like) are used to calculate the review informative score S(I). The proposed review informative score S(I) consists of two parts: content score S(C) and popularity score S(P). S(C) is calculated using the product metadata and with the help of SentiWordNet whereas S(P) is calculated using the review posted date and its helpful votes. S(I) is the sum of both S(C) and S(P) multiplied by weights, where the sum of two weights is one. The dataset is then split into two sets, labelled set L and unlabelled set U, and then transformed into vectors. The labelled set L is used to train a model, then the model is used to predict labels for unlabelled set U. The predicted labels are either positive or negative. Along with the confidence given by the model, the pseudo-labelled set is ranked using the sum of both confidences and review informative score S(I) multiplied with weights, where the sum of two weights is one. Confidence is a value representing the confidence level of the model predicting the label correctly. The top ten per cent of predicted instances are chosen for data augmentation and dropped from the unlabelled set U. The model is then retrained using the augmented data and tested using a test set to measure the accuracy of the model. If accuracy satisfies the expectation, a final model is produced. If not, back to the prediction step and continue to retrain the model. The flow of the proposed methodology is shown in Fig. 1.

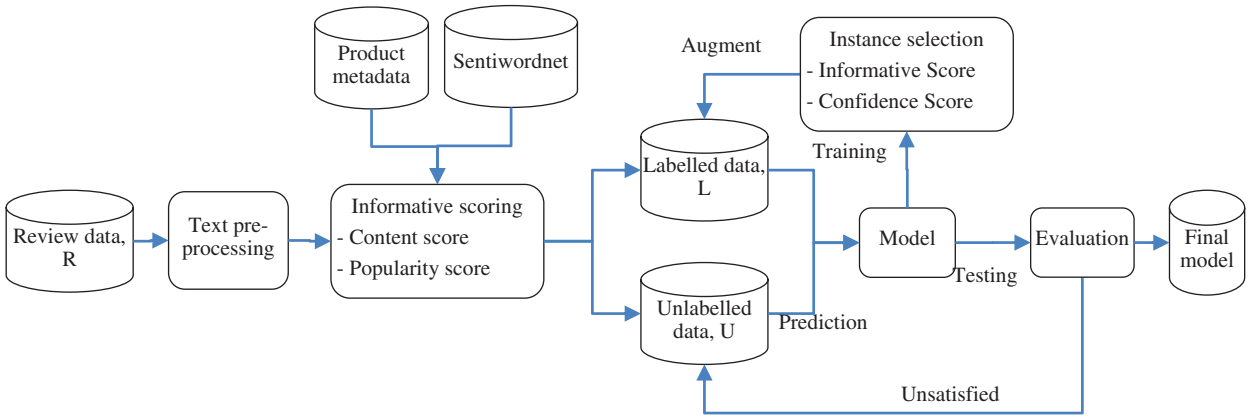

Figure 1: Semi-supervised sentiment classification with review informative score for instance selection

3.2.1 Text Preprocessing

In this study, we used the publicly available dataset, Amazon review data by [20]. The dataset is first processed using the text preprocessing steps as follows.

i. Remove duplicates: Duplicated reviews are removed to improve the quality of training data for better model accuracy. Duplicates might have caused the model to produce a biased result.
ii. Rating conversion: The ratings in review data are converted into two classes, positive and negative. The rating of review data is in a range of 1 to 5, of which 5 is the highest rating and 1 is the lowest rating. Hence, reviews with ratings 4 and 5 are labelled as positive, whereas reviews with ratings 1 and 2 are labelled as negative. Reviews with a rating of 3 are considered neutral and used in this study. In other situations, one may omit this step or adjust the conversion rules accordingly.

iii. Sentence boundary disambiguation: This task is to deconstruct a piece of text into sentences. The review text is segmented into sentences by recognizing the end of a sentence using end-of-sentence punctuation marks, such as period ".", question mark "?", and exclamation mark "!".

iv. Part-of-speech (POS) tagging: Part-of-speech (POS) tagging, also known as grammatical tagging, is a process of categorizing and assigning each word in the text a specific tag in correspondence with its part of speech, based on the definition and context of the word. POS tags describe the syntactic category of a word within a sentence or text; for example, noun, verb, and adjective.

v. Stopwords removal: Stopwords are the words that appear frequently in the text that do not provide significant information about the sentiment polarity of a given text. For example, "the", "a" and "an" are words that appear frequently in English text. However, removing all the stopwords is not always the best choice. Thus in this study, only stopwords in the following POS tag categories are removed: DET, NUM, PUNCT, SYM, X (other) and SPACE.

vi. Dependency parsing: Dependency is a binary relationship that elucidates the relationships between the words. This process is to analyse the grammatical structure and determine the dependency tag that represents the relationship between two words in a sentence. For example, in the sentence "it stayed in pocket", there is a nominal subject relationship nsubj(stayed, it) between the words "it" and "stayed". The outcome of this process is a dependency tree that depicts the grammatical relations existing among different words. The nodes in a dependency tree mark the syntactical class of each word, whereas the edges label the ordered structure of grammatical relations between words (see Fig. 2).

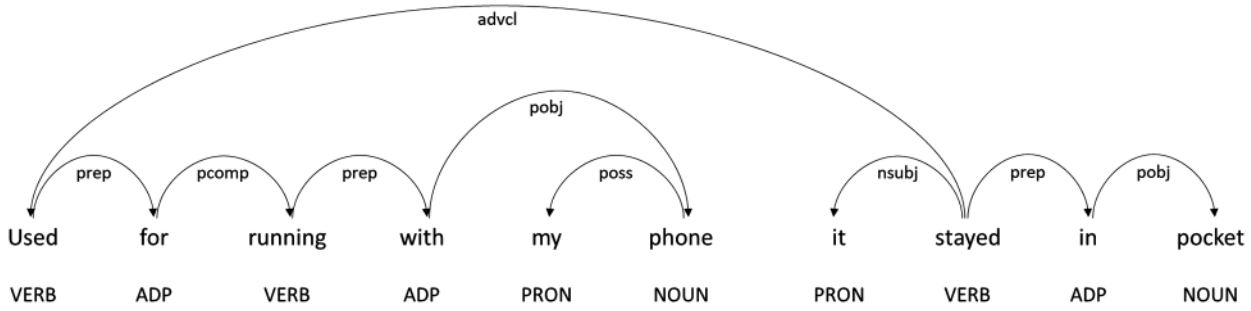

Figure 2: An example outcome of dependency parsing

A snippet of the product review dataset after preprocessing is shown in Table 3. The index column is the sentence ID, whereas the reviewID column is the review ID. If a review has more than one sentence, the value of reviewID will remain the same and the Index value will be increased. For example, consider this review, "These are great but not much better then gen1. Only addition is Siri feature. I will rather buy the previous model on discount and save some green."

ReviewID will remain the same as "0" for these 3 sentences, whereas the Index value for each new sentence will be "0, 1, 2". The reviewPos column is the sentence tagged with POS tag and the reviewDep column is the sentence parsed with dependency.

Table 3: A snippet of product review dataset after preprocessing

Atrributes	Values
Index	0
reviewID	0
asin	B0829DCVN6
vote	22
overall	1
reviewTime	2009-05-17
reviewPos	[('nice', 'ADJ'), ('set', 'NOUN',), ('of', 'ADP'), ('cutting', 'VERB'), ('boards', 'NOUNS')]
reviewDep	['amod(set-NOUN, nice-ADJ)', 'ROOT(set-NOUN, set-NOUN)', 'prep(set-NOUN, of-ADP)', 'compound(boards-NOUN, cutting-NOUN)', 'pobj(of-ADP, boards-NOUN)']

3.2.2 Review Scoring

A processed text is matched with a product ID to retrieve its product metadata such as product title, product features and product description. Then, a list of nouns is extracted from the product metadata. When the noun from a review is matched with the list, adjectives, verbs and adverbs related to the noun in review will be checked against SentiWordNet. In SentiWordNet, a word is considered neutral if the positive and negative scores of the word are the same. For instance, "handy" is a neutral word as it has the same positive and negative scores, which is 0.125. Whereas the word "able" has a positive score of 0.125 and a negative score of 0, it is considered a positive word. Therefore, SentiWordNet is used to check whether the adjectives, verbs or adverbs are not neutral.

Besides, posted date and helpful votes of the review are also utilized to calculate the review informative score S(I). After the S(I) of each review is calculated, the S(I) of all data points is normalized into a range of 0 and 1 using min-max normalization in order not to heavily influence the ranking of the unlabelled set. The concept and formula of the review informative score S(I) will be introduced in the following Section 3.2.2(a). Data are transformed into vectors and the labelled set is used to train the model. Then, the model is asked to predict the labels for the unlabelled set. Prediction confidence, a value in the range of 0 and 1, is assigned to each unlabelled review along with the label predicted by the model.

(a) Review informative score S(I)

An opinion is defined by [21] as a quintuple, (e, a, s, h, t), where e is an entity, a is an aspect of e, s is the sentiment on aspect a, h is the opinion holder, and t is the time when opinion is expressed by h. Following this definition, four components are included in review informative scoring except for the opinion holder h. A product review is informative when it expresses sentiments on many aspects of the product and is posted recently. Older reviews' information might be outdated as sellers may improve their products and services over time. Besides the user-to-product interactions such as review text and review posted date, user-to-user interactions such as helpful votes are also taken into consideration in review informative scoring. When a potential buyer finds the reviewer's review helpful, one may upvote the review or no action is taken. A product review is informative when it has many helpful votes from potential buyers. Review informative score S(I) consists of two parts; content score S(C)

and popularity score S(P); w_1 and w_2 are weights assigned to the scores.

$$S(I) = w_1 S(C) + w_2 S(P), \ \textit{where } w_1 + w_2 = 1 \tag{1}$$

(b) Content score S(C)

The purpose of the content score is to represent the quality of the review and its relevance to the particular product. To determine the content score of the review, product metadata such as product ID, product title, product features and product description are utilized in the process. The product ID is used to match a review to a particular product as a product can have multiple reviews. Product title, product features and product description are used to determine the item posted for sale and its aspects. Product title is a required attribute, but product features and product description may be optional attributes and they are used to describe the aspects of the products. Thus, both attributes are included in the content score to increase the possibility of having the aspects of products captured in the review.

For example, an item with the product title "Fuji Red Apple", its product description says "Crispy and fragrant Fuji red apple, 5 pieces wrapped in a packet" and the product features listed "Juicy and bursting in flavour", "High in fibre" and "Imported from Japan". The nouns found in the product title, product description and product features are "apple", "packet", "flavour", "fibre" and "Japan". By comparing the nouns found in product metadata and product review, the adjectives, verbs and adverbs in the review that have a dependency on the nouns found in both metadata and review are determined. Then, positive and negative scores of adjectives, verbs and adverbs are checked in the SentiWordNet. If the adjective, verb or adverb has a different positive score and negative score, meaning there is a sentiment expressed towards one of the aspects of the product, thus adding one to the content score S(C). Fig. 3 shows the pseudo-code used in determining the content score S(C). The final S(C) is in the range of zero to infinity, and the maximum value could not be identified because the number of sentiments expressed in the review is not limited. Besides, we would like to avoid the content score S(C) taking a bigger part in Eq. (1). Therefore, the value is normalized into the range of 0 and 1 using min-max normalization to reduce its influence in calculating the review informative score S(I).

(c) Popularity score $S(P)$

Popularity score S(P) is composed of two parts, the number of helpful votes and the period of a review posted. Its purpose is to represent the value of a review and its relation to the particular product. The higher the helpful vote, the more value the review holds. However, the value of the review has deteriorated over time as it may not hold anymore because sellers may improve their product based on feedback. Besides, whenever a review is old, it becomes more vulnerable to being read by other users and the number of helpful votes increases. Thus, popularity score S(P) is written as

$$S(P) = \frac{\textit{number of helpful votes}}{\textit{period of posted review}} = \frac{\textit{number of helpful votes}}{(\textit{current date–posted date})} \tag{2}$$

For example, review A posted on 2019/8/7 has 51 helpful votes; this review is shown on top of the review section in default (top reviews view). When a potential customer checks the review section of this product, the first review read is the top review of this product. If that potential buyer finds the review helpful, he may upvote the review. But recent review B posted on 2020/6/5 has the same number of helpful votes as review A is a better review. A comparison and calculation between review A and review B are depicted in Fig. 4.

```
Input: review, productID, productTitle, productFeature, productDesc
Output: content score, S(C)
begin
   foreach review do
      if productID in review == productID in metadata
          retrieve productTitle, productFeature and productDesc
          get list of noun
          foreach noun in list == noun in reviewText do
             get list of adjective, verb and adverb that has dependency with the noun
             foreach adjective, verb or adverb in review found do
                check sentiwordnet score table
                if PosScore(adjective, verb or adverb) != NegScore(adjective, verb or adverb)
                    add 1 to content score, S(C)
                end
             end
          end
       end
   end
end
```

Figure 3: Pseudocode for determining the content score

Review A:	Review B:
vote = 51	vote = 51
currTime = (2021,07,16)	currTime = (2021,07,16)
reviewTime = (2019,08,07)	reviewTime = (2020,06,05)
periodPosted = currTime – reviewTime = 709 days	periodPosted = currTime – reviewTime = 406 days
S(P) = vote/periodPosted = 0.0719	S(P) = vote/periodPosted = 0.1256

Figure 4: A comparison and calculation of example review A and B

(d) Instance selection based on confidence and informative score

The unlabelled set U is then ranked using the following equation:

$$x_1 confidence + x_2 S(I), \text{ where } x_1 + x_2 = 1 \tag{3}$$

Both x_1 and x_2 were set as 0.5 in this study to achieve a balanced influence from both confidence and review informative score S(I) for the ranking. Confidence is a value returned by the model for each prediction, showing the probability of the input falling into different classes. Different algorithms have different equations for calculating confidence, for example, the formula for Naïve Bayes is written as $P(A|B) = (P(B|A) \times P(A))/P(B)$. The top ten per cent of predicted instances in the ranking are selected for data augmentation, included in the labelled set and used for the next round of model training. Selected instances are then dropped from the unlabelled set. After the model is retrained using the augmented input, the model is tested using the test set and the accuracy of the model is measured. Accuracy is defined as the percentage of correct predictions for test data (see Eq. (4)).

$$Accuracy = \frac{N_{correct}}{N_{total}} \times 100\% \quad (4)$$

$N_{correct}$ is the number of reviews predicted with correct polarity and N_{total} is the total number of reviews tested.

If the accuracy of the model does not reach a satisfactory result, repeat the prediction step to get predictions from the retrained model for the unlabelled set and continue to retrain the model. A final model is produced when the accuracy of the model reaches a satisfactory result.

4 Evaluations

4.1 Dataset Description

The dataset used in this experiment is the Amazon review dataset [20]. It consists of Amazon product reviews in 29 domains such as books, electronics, toys and games, having a total number of 233.1 million product reviews. The dataset also includes product information that has been reviewed as metadata. 5 out of 29 domains that had the highest number of reviews were chosen to include in this experiment. The top 5 domains are books, clothing, shoes and jewellery, home and kitchen, electronics, sports and outdoors. Each review in the Amazon review dataset is submitted with a star rating range of 1 to 5. The rating is used as a reference for overall polarity, which is either positive or negative at the document level. The summary of the dataset is shown in Table 4. Reviews with star ratings 4 and 5 are translated into positive reviews, whereas reviews with star ratings 1 and 2 are translated into negative reviews. Reviews with a neutral sentiment that have a star rating of 3 and duplicated reviews are grouped as others and they are excluded from the evaluation.

Table 4: Summary of total reviews by polarity

Domain	Total	Positive	Negative	Others*
Books	51 331 621	38 561 230 (75.12%)	12 704 051 (24.75%)	66 340 (0.13%)
Clothing, shoes and jewellery	32 292 099	25 613 581 (79.32%)	6 628 193 (20.53%)	50 325 (0.15%)
Home and Kitchen	21 928 568	13 856 240 (63.19%)	8 041 736 (36.67%)	30 592 (0.14%)
Electronics	20 994 353	16 352 069 (77.89%)	4 579 745 (21.81%)	62 539 (0.30%)
Sports and outdoors	12 980 837	8 362 595 (64.42%)	4 571 947 (35.22%)	46 295 (0.36%)

Note: *Others included neutral reviews with a star rating of 3 and duplicated reviews

Each domain dataset is made up of two datasets, review and metadata. The review contains the records of product reviews, whereas metadata contains the records of product information in a particular domain. Attributes of review and metadata which contain information that is not in the form of text are dropped in the experiment. Following the concept of review informative score S(I), *vote*, *reviewText*, *overall and reviewTime* from the review are kept for the sentiment classification. As for the metadata, only the *title, feature and description* are used to identify entities and features of entities in review text for the calculation of content score S(C).

4.2 Evaluating Informative Scoring

The first research question is to develop a scoring formula for calculating informative levels of reviews and improve the performance of semi-supervised sentiment classification. Thus, the first

experiment goal is to find out whether a semi-supervised model that incorporated a review informative score is performing better than a semi-supervised model without including a review informative score in its training. Besides, we also would like to check whether a semi-supervised model incorporating a review informative score produces comparable results with a supervised model. For evaluation, we have trained sentiment classification models using 3 methods, 4 algorithms and reviews from 5 domains in Amazon review dataset. The first method is the fully supervised method which acts as the baseline, the second method is the semi-supervised method without the introduction of an informative score, and finally, the third method is the semi-supervised method incorporating an informative score. The 4 algorithms used in this experiment include 2 conventional machine learning algorithms and 2 deep learning algorithms: NB, SVM, CNN and RNN. The 5 chosen domains in the Amazon review dataset are books, clothing, shoes and jewellery, home and kitchen, electronics, sports and outdoors. Experiments were performed on a 3.2 GHz i7 machine with 32 GB RAM and Nvidia GeForce RTX 2060 on Win10.

We compared the performances of the supervised sentiment classification models, semi-supervised sentiment classification models without incorporating review informative score S(I) and the semi-supervised sentiment classification models incorporating review informative score S(I). Weights of w_1 and w_2 in Eq. (1), and x_1 and x_2 in Eq. (3) are all set as 0.5 in this experiment. The results are measured using model accuracy. In the training of sentiment classification models, reviews of each domain are split into the training set and test set in the ratio of 8:2. Supervised models require fully labelled data, thus the training set is fully labelled in the baseline training. Whereas the semi-supervised model requires only partially labelled data for training. Considering the 8-part supervised training set as a full training set, the training set is further split into a labelled set and an unlabelled set with the ratio of 4:6. The unlabelled training subsets are used in acquiring model predictions, and sentiment labels of data points. A breakdown of the dataset is shown in Fig. 5.

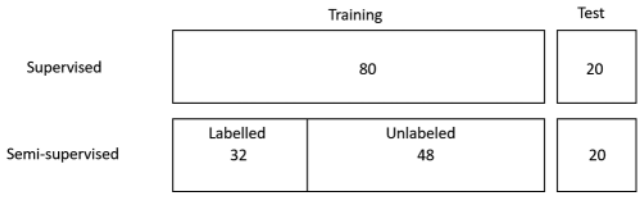

Figure 5: Partitioning of datasets

Secondly, we aim to investigate the optimum ratio of labelled and unlabelled data in semi-supervised learning. Therefore, in the second experiment, three ratios of labelled and unlabelled data for semi-supervised sentiment classification were compared. The three ratios of labelled and unlabelled data chosen were 20:80, 30:70 and 40:60 in this experiment. The ratio of 10:90 has a higher probability that the semi-supervised model will not perform well, thus being excluded from the experiment. Whereas for a percentage of labelled data that is equal and higher than 50, this defeats the purpose of semi-supervised learning and is not practical. Therefore, the ratios 50:50, 60:40, etc are ruled out in the experiment. Following the third objective, we have further evaluated the semi-supervised CNN in different dataset ratio settings using the clothing, shoes and jewellery dataset because this combination performed the best and achieved accuracies above 80% in both semi-supervised settings.

5 Results

5.1 Results of Semi-Supervised Learning Using Informative Score

In this section, we present the results of our proposed approach, semi-supervised sentiment classification incorporating review informative score on 5 domains from the Amazon dataset. Each domain dataset is compared using a combination of 3 methods and 4 algorithms. The 3 methods are supervised, semi-supervised without an informative score, and semi-supervised with an informative score, whereas the 4 algorithms are NB, SVM, CNN and RNN. A total of 12 models were compared for each domain dataset. Table 5 show the model accuracies for the 5 different domains mentioned.

Table 5: Results comparing supervised, semi-supervised without and with an informative score

Domain	Models	Supervised (%)	Semi-supervised without informative score (%)	Semi-supervised with informative score (%)
Book	NB	86.23	79.36	84.62
	SVM	85.64	**80.63***	**86.65***
	CNN	**87.26***	78.35	82.51
	RNN	86.51	78.62	82.35
Clothing, shoes and jewellery	NB	82.36	75.45	80.62
	SVM	84.36	77.55	81.34
	CNN	85.62	**81.52***	**87.13***
	RNN	**84.67***	79.69	86.51
Home and kitchen	NB	77.65	73.62	76.51
	SVM	81.63	**79.51***	82.48
	CNN	82.45	77.32	82.50
	RNN	**84.62***	78.05	**83.62***
Electronics	NB	83.62	78.52	81.26
	SVM	79.35	76.35	81.45
	CNN	**84.31***	**78.62***	**83.54***
	RNN	78.62	73.64	80.24
Sports and outdoor	NB	83.62	**81.62***	**84.35***
	SVM	77.33	73.65	79.62
	CNN	**84.15***	77.66	82.62
	RNN	82.61	78.54	83.52

Note: *Best algorithm for each method per domain.

It is observed that the accuracy of supervised models is better than semi-supervised models without an informative score. However, their difference in accuracy is small, in the range of 2 to 9 per cent. For example, for the accuracies of CNN for the book's domain, the supervised model has an accuracy of 87.26%, whereas the semi-supervised model without an informative score has an accuracy of 78.35%. The difference is calculated by operating subtraction between the two accuracies, which is 8.91%, rounded up to 9%. But considering only 40 per cent of the labelled data used in training

semi-supervised models, we can say that the performance of semi-supervised models is comparable to the supervised models that use fully labelled data. On the other hand, all the semi-supervised models with an informative score performed better compared to semi-supervised models without an informative score. This indicates that the informative score has a positive effect on model accuracy. Although the accuracy did not surpass supervised models, the accuracy of semi-supervised models with an informative score is considered good as they are close to the accuracy of supervised models that require fully labelled data.

The highest accuracy for each method is bold in the table, showing the best-performed method for each algorithm. From the row showing the best algorithm for each method, we can see that the results shown by deep learning are better than conventional machine learning algorithms; 10 out of 15 of them are deep learning algorithms. Even when compared with the supervised model, deep learning is still performing better than conventional machine learning algorithms. This is expected and aligns with the ability of deep learning algorithms, as the size of the dataset used is bigger than the conventional machine learning algorithms can handle well. Among the chosen domains, the book dataset produced better results in the range of 78.35% to 87.26%. We have observed that the product descriptions and reviews in the book's dataset tend to be longer than in the other datasets. Besides, reviews tend to comment on books' contents, which usually can be found in product descriptions and produce a matching noun for the content score. These result in an increased occurrence of high content score S(C).

5.2 Results of Label vs. Unlabeled Data Ratio and Instance Selection Strategy

The second part of the experiment further explored different ranking settings and weights. In Table 6, 40 per cent of labelled data used in semi-supervised training produced a model with an accuracy of 86.78%. The results showed that higher ratios of labelled data in semi-supervised training yielded higher model accuracies. If the ratio of labelled data goes above 50 per cent, the experiment defeats the purpose of semi-supervised learning, which is targeting to lower the number of labelled data needed close to none, enabling the power of unsupervised learning.

Table 6: Results of semi-supervised CNN with instance selection based on review informative score $S(I)$

Ratio|Weight	0.5:0.5	0.6:0.4	0.4:0.6
20:80	**80.72**	80.62	80.65
30:70	84.28	**84.89**	84.18
40:60	**86.71**	86.52	85.64

Note: Weight, w_1: w_2 of S(I) (content: popularity); Ratio, labeled: unlabeled.

In Table 7, we focused on the weights of S(I) and found balanced weights (content/popularity) of 0.5/0.5 gives the best results. We observed from our experiment that content score S(C) heavily depends on product details given and wording in reviews as it is calculated based on the matching nouns from both. There are a variety of word choices that can be used to describe the same thing, for example, phone, handphone and sometimes the words "it" or "they" are used to refer to the product. Thus sentiments in the reviews may not be fully captured. Whereas popularity score S(P) relies on the helpful votes from potential buyers and not all the reviews have received a helpful vote. The helpful vote reflects how buyers perceive the review as informative in their buying decision. Therefore, a balance between

these two factors of review informative score S(I) produced the best results. The experiments also focused on different weights of instance selection based on Eq. (3) and review informative score S(I). Comparing across the three tables, the model yields the best accuracies when both the instance selection weights based on Eq. (1) (content/popularity) and weights for Eq. (3) (confidence/informative) have a balanced value of 0.5/0.5. The confidence score is determined by the prediction model and its label quality. For example, the label quality of a pseudo-labelled sample in SVM is higher when the sample is further away from its hyperplane h. This means SVM is in favour of the distribution of samples that are further away from h. On the other hand, review informative score S(I) is based on opinion towards the product in review and the support from potential buyers. Having a balanced amount of support evidence from two factors helped to select the impactful unlabeled instances with predictions for model training.

Table 7: Results of semi-supervised CNN with instance selection based on confidence and review informative score $S(I)$

Ratio	Weight			
	Instance selection (confidence:S(I))	S(I) (content:popularity)		
		0.5:0.5	0.6:0.4	0.4:0.6
20:80	**0.5:0.5**	**87.13**	81.32	85.64
30:70		**88.62**	86.84	84.31
40:60		**89.52**	83.32	84.82
20:80	**0.6:0.4**	**86.43**	79.81	80.12
30:70		**86.34**	81.32	82.16
40:60		**87.34**	84.54	85.91
20:80	**0.4:0.6**	**84.48**	80.23	82.89
30:70		**85.52**	82.51	80.91
40:60		**86.42**	83.21	84.35

Note: Weight, w_1: w_2 of S(I) (content: popularity); Ratio, labeled: unlabeled; Instance selection: confidence:$S(I)$.

6 Conclusion

This paper proposed a review's informative score S(I) and its implementation in the pseudo-labelled instances selection process of semi-supervised training. Current works have mostly focused on confidence score when selecting the pseudo-labelled data points to be included in the subsequent training cycle; S(I) can be served as an alternative alongside confidence to select informative pseudo-labelled instances that have positive impacts for the next training. S(I) can be generalized to different domains as well. Although its performance is lacking when compared to a supervised model, S(I) can improve the accuracies of the semi-supervised models. Optimum weights of S(I) and instance selection were also determined in the experiment. In conclusion, the main contribution of this research is the presentation of S(I) to select informative pseudo-labelled instances for semi-supervised training cycles. Most of the existing work focused solely on confidence and the threshold value is vague, mentioning only high and low without an actual number. Our attempt linked how humans perceive an opinion as

informative to instance selection strategy in semi-supervised learning. The results showed that semi-supervised training with only a small amount of labelled data could produce models with comparable results to the supervised models.

6.1 Limitations

Some limitations are observed for future work of this research. First, S(I) relies on product description and product features to determine the entity or attributes of the entity mentioned in the review text. If the seller did not upload a description of the product or product features, then the method could not detect the entity or attributes of the entity mentioned in the review text and failed to capture the sentiment expressed. Next, the method also does not consider subject pronouns in a sentence. This may cause not all the sentiments expressed towards the product in the review to have been captured. Look at the following example of a review: "*First I have to say, you will get a good whooping from these when you first try to find your rhythm. I messed up so many times but I can do it better now, I like them because it helps me warm up quicker than stretching before exercise. You start to feel it in your legs, and realize you ain't a kid anymore*." This review did not mention what the product bought in the review, however, the user uses "them" and "it" to refer to the product bought. The method failed to capture sentiments expressed in the review and gave an inaccurate review informative score. Lastly, the results of the classification can be strenghtened with statistical test to support the outcome of the evaluations.

Funding Statement: This research is supported by Fundamental Research Grant Scheme (FRGS), Ministry of Education Malaysia (MOE) under the project code, FRGS/1/2018/ICT02/USM/02/9 titled, Automated Big Data Annotation for Training Semi-Supervised Deep Learning Model in Sentiment Classification.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[2] K. Dave, S. Lawrence and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, " in *Proc. WWW '03*, Budapest Hungary, pp. 519–528, 2003.

[3] E. Cambria, S. Poria, A. Gelbukh and M. Thelwall, "Sentiment analysis is a big suitcase," *IEEE Intelligent Systems*, vol. 32, no. 6, pp. 74–80, 2017.

[4] V. Iosifidis and E. Ntoutsi, "Large scale sentiment learning with limited labels," in *Proc. KDD '17*, Halifax NS Canada, pp. 1823–1832, 2017.

[5] J. Levatić, M. Ceci, D. Kocev and S. Džeroski, "Semi-supervised classification trees," *Journal of Intelligent Information Systems*, vol. 49, no. 3, pp. 461–486, 2017.

[6] W. Zhang, X. Tang and T. Yoshida, "TESC: An approach to text classification using semi-supervised clustering," *Knowledge-Based Systems*, vol. 75, pp. 152–160, 2015.

[7] W. He, X. Huang, G. Tsechpenakis, D. Metaxas and C. Neidle, "Discovery of informative unlabelled data for improved learning, " in *Proc. IEEE ICCV Int. Workshop on Modeling People and Human Interaction*, Beijing, China, 2005.

[8] Q. Tian, J. Yu, Q. Xue and N. Sebe, "A new analysis of the value of unlabelled data in semi-supervised learning for image retrieval, " in *Proc. ICME 2004*, Taipei, Taiwan, pp. 1019–1022, 2004.

[9] M. Fernández-Gavilanes, T. Álvarez-López, J. Juncal-Martínez, E. Costa-Montenegro and F. J. González-Castaño, "Unsupervised method for sentiment analysis in online texts," *Expert Systems with Applications*, vol. 58, no. 1, pp. 57–75, 2016.
[10] X. Hu, J. Tang, H. Gao and H. Liu, "Unsupervised sentiment analysis with emotional signals, " in *Proc. WWW '13, Association for Computing Machinery*, New York, NY, USA, pp. 607–618, 2013.
[11] Y. Zhou, M. Kantarcioglu and B. Thuraisingham, "Self-training with selection-by-rejection, " in *Proc. IEEE 12th Int. Conf. on Data Mining*, Brussels, Belgium, pp. 795–803, 2012.
[12] Y. Han, Y. Liu and Z. Jin, "Sentiment analysis via semi-supervised learning: A model based on dynamic threshold and multi-classifiers," *Neural Computing and Applications*, vol. 32, pp. 5117–5129, 2020.
[13] N. F. F. da Silva, L. F. S. Coletta, E. R. Hruschka and J. Hruschka, "Using unsupervised information to improve semi-supervised tweet sentiment classification," *Information Sciences*, vol. 355–356, pp. 348–365, 2016.
[14] F. H. Khan, U. Qamar and S. Bashir, "A Semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet," *Knowledge and Information Systems*, vol. 51, no. 3, pp. 851–872, 2016.
[15] S. Lee and W. Kim, "Sentiment labeling for extending initial labelled data to improve semi-supervised sentiment classification," *Electronic Commerce Research and Applications*, vol. 26, no. C, pp. 35–49, 2017.
[16] J. Chen, J. Feng, X. Sun and Y. Liu, "Co-training semi-supervised deep learning for sentiment classification of MOOC forum posts," *Symmetry*, vol. 12, no. 1:8, pp. 1–24, 2020.
[17] K. Bijari, H. Zare, E. Kebriaei and H. Veisi, "Leveraging deep graph-based text representations for sentiment polarity applications," *Expert Systems with Applications*, vol. 144, no. 15, pp. 1–10, 2020.
[18] J. Duan, B. Luo and J. Zeng, "Semi-supervised learning with generative model for sentiment classification of stock messages," *Expert Systems with Applications*, vol. 158, no. 15, pp. 1–9, 2020.
[19] M. Sharma and M. Bilgic, "Evidence-based uncertainty sampling for active learning," *Data Mining and Knowledge Discovery*, vol. 31, no. 1, pp. 164–202, 2017.
[20] J. Ni, J. Li and J. McAuley, "Justifying recommendations using distantly-labelled reviews and fined-grained aspects, " in *Proc. EMNLP-IJCNLP*, Hong Kong, China, ACM, pp. 188–197, 2019.
[21] B. Liu, "Document Sentiment Classification," in *Sentiment Analysis and Opinion Mining*, 1st ed., Cham: Springer, pp. 23–36, 2012.