Negative Confidence-Aware Weakly Supervised Binary Classification for Effective Review Helpfulness Classification

Xi Wang
University of Glasgow, UK
x.wang.6@research.gla.ac.uk

Iadh Ounis
University of Glasgow, UK
iadh.ounis@glasgow.ac.uk

Craig Macdonald
University of Glasgow, UK
craig.macdonald@glasgow.ac.uk

ABSTRACT

The incompleteness of positive labels and the presence of many unlabelled instances are common problems in binary classification applications such as in review helpfulness classification. Various studies from the classification literature consider all unlabelled instances as negative examples. However, a classification model that learns to classify binary instances with incomplete positive labels while assuming all unlabelled data to be negative examples will often generate a biased classifier. In this work, we propose a novel Negative Confidence-aware Weakly Supervised approach (NCWS), which customises a binary classification loss function by discriminating the unlabelled examples with different negative confidences during the classifier's training. NCWS allows to effectively, unbiasedly identify and separate positive and negative instances after its integration into various binary classifiers from the literature, including SVM, CNN and BERT-based classifiers. We use the review helpfulness classification as a test case for examining the effectiveness of our NCWS approach. We thoroughly evaluate NCWS by using three different datasets, namely one from Yelp (venue reviews), and two from Amazon (Kindle and Electronics reviews). Our results show that NCWS outperforms strong baselines from the literature including an existing SVM-based approach (i.e. SVM-P), the positive and unlabelled learning-based approach (i.e. C-PU) and the positive confidence-based approach (i.e. P-conf) in addressing the classifier's bias problem. Moreover, we further examine the effectiveness of NCWS by using its classified helpful reviews in a state-of-the-art review-based venue recommendation model (i.e. DeepCoNN) and demonstrate the benefits of using NCWS in enhancing venue recommendation effectiveness in comparison to the baselines.

ACM Reference Format:
Xi Wang, Iadh Ounis, and Craig Macdonald. 2020. Negative Confidence-Aware Weakly Supervised Binary Classification for Effective Review Helpfulness Classification. In *The 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340531.3411978

1 INTRODUCTION

A generic binary classifier focuses on modelling data with both positive and negative ground truth labels. However, in many binary classification applications, such as review helpfulness classification or relevant document classification, it is common for the ground truth to contain only a few positive and many unlabelled instances. In particular, the unlabelled instances contain both positive and negative instances. For example, in online book reviews, only a few reviews are deemed helpful by other users (positive instances) and many reviews are neither assessed helpful nor unhelpful by the users (unlabelled instances). This data incompleteness harms the accuracy of identifying positive and negative instances [12]. In this paper, we address the problem of classification with incomplete positive and abundant unlabelled instances. On the other hand, weakly supervised classifiers have been proposed to address classification with noisy, limited or imprecise data resources [46]. Indeed, binary classification with incomplete positive instances and unlabelled instances can also be addressed by a weakly supervised learning process [2].

Among various weakly supervised approaches, the Positive-Unlabelled learning approach (aka PU learning) is a popular solution in addressing cases where the data has few positive instances and many unlabelled instances, by leveraging estimates of the class priors [3, 35]. However, as du Plessis [12] argued, class prior estimation-based solutions lead to a systematic estimation bias. Therefore, we propose to conduct binary weakly supervised classification on data with incomplete positive instances and unlabelled instances without the aid of an estimated class prior for the unlabelled examples. Moreover, du Plessis et al. [12] proposed to address the classifier bias problem of PU learning by applying different loss functions for the positive and unlabelled classes (an approach that we will refer to as **C-PU**). We also argue that using two customised loss functions for the positive and unlabelled data could help a classification model to fully leverage labelled and weakly labelled data, respectively. Hence, we follow du Plessis et al. [12] by using different loss functions to different classes. Meanwhile, Ishida et al. [19] proposed a classification approach that solely relies on the positively-labelled instances to generate a classifier according to the positivity or the confidence of the examples being positive. However, this approach filters out unlabelled examples, which can lead to a problem of information loss and therefore negatively impacts the classification performance. Inspired by earlier works on PU learning [12, 19], we propose a negative confidence-aware weakly-supervised binary classification approach, **NCWS**, which considers both positive and unlabelled examples with corresponding customised loss functions. In particular, we use an ordinary loss function for the positive class instead of the composite loss function $l(z) - l(-z)$ used in [12]. Our approach is also different from [19], which only uses the positive class for binary classification. For the negative class, instead of using an ordinary loss function $l(z)$ for the unlabelled class, we design a customised loss function for the unlabelled class by considering the distinct probabilities of the unlabelled instances to be negative (i.e. negative confidence). We leverage the properties of the unlabelled instances (e.g. their age in the dataset since their creation time) to estimate their probabilities

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3411978

of being negative. As we will explain later, these properties are chosen based on their likely correlation with the negative class. Indeed, our proposed NCWS approach uses additional complementary information to the content and labels of instances (namely, the aforementioned properties of instances) to infer the likelihood of the unlabelled examples belonging to the negative class.

To evaluate the effectiveness of our proposed NCWS approach, we apply NCWS to address the user review helpfulness classification task. User reviews contain various types of user opinions, including user preferences, item reputations, and item properties. However, low-quality reviews bring a certain inconvenience to users when assessing reviews and making decisions on buying a product, visiting a venue or watching a movie. It is also beneficial to business owners to be able to identify the helpfulness of reviews for their products. For example, identifying helpful reviews allows to selectively present reviews to customers, thereby supporting them in making informed decisions [31]. To the best of our knowledge, most review helpfulness classification studies in the literature use a binary classification setup and consider reviews receiving sufficient votes[1] as positive and the rest of reviews as negative [8, 22]. However, reviews without helpful labels might have not yet gained views, or might have been hidden from users by the interface. Consequently, we often observe a bias towards reviews that have been presented to users. Indeed, the unlabelled reviews could still be either helpful or unhelpful. Therefore, due to the prevalence of unlabelled instances in the review helpfulness classification task and the associated challenges, we argue that this task is a good scenario for examining the performance of our NCWS approach in addressing the weakly supervised binary classification problem.

In the review helpfulness classification task, modelling and representing reviews is key to the development of effective review helpfulness classifiers. Most existing approaches model the content of user reviews and the corresponding rating information, then make predictions on the review helpfulness labels (i.e. helpful or unhelpful). In this paper, we reproduce many state-of-the-art review helpfulness classification approaches as baselines and refer to these approaches as the *basic* classifiers. We select the basic classifiers with the best performances and use them to evaluate the effectiveness of our NCWS approach. We then thoroughly examine the performance of our NCWS approach in identifying further helpful reviews. We extensively validate the effectiveness of our proposed NCWS approach in review helpfulness classification by investigating the extent to which accurately identifying additional helpful reviews can further enhance an existing review-based venue recommendation model. In particular, we use the DeepCoNN model proposed by Zheng et al. [45], which is a popular state-of-the-art review-based recommendation model.

The main contributions of this paper are summarised as follows:
• We propose a weakly supervised binary classification correction approach (NCWS), which leverages positive unlabelled learning and uses a negative confidence-based loss function for modelling the unlabelled examples.
• We show how to integrate NCWS within different popular binary classifiers, including SVM, CNN and BERT-based classifiers [10], to effectively address the review helpfulness classification task.
• We evaluate the effectiveness of our proposed NCWS approach by comparing it with several existing state-of-the-art approaches in the literature, including the SVM Penalty-based approach (SVM-P), a PU learning-based approach (C-PU) and a positive confidence-based approach (P-conf).
• We evaluate the performance of NCWS on three real-world datasets, namely one from Yelp (Venue reviews) and two from Amazon (Kindle and Electronics reviews).
• We validate the utility of NCWS, by using its predicted helpful reviews as input to a state-of-the-art review-based recommendation model (i.e. DeepCoNN).

The paper is organised as follows. In Section 2, we describe related work. We state the tackled problem and the methodology underpinning our NCWS approach in Section 3. In Section 4, we list a number of review helpfulness classifiers. Next, in Section 5, we introduce our research questions and our three used Yelp and Amazon-based datasets. We also describe the experimental setup for the basic classifiers, the baseline approaches and our NCWS approach, as well as the used evaluation metrics for performance comparison. In Section 6, we analyse the results from the experiments to answer the research questions. In Section 7, we demonstrate the utility of our NCWS approach by improving the accuracy of a venue recommendation approach using the accurately classified reviews. Finally, we provide concluding remarks in Section 8.

2 RELATED WORK

In this section, we describe related approaches in weakly-supervised learning and review helpfulness classification.

2.1 Weakly-supervised Approaches

Many studies from the classification literature typically consider unlabelled instances as negative examples [16, 25, 27]. However, the simple grouping of unlabelled examples into a single negative class leads to inaccurate and biased binary classifiers [12, 23]. Such an inaccuracy comes from the unlabelled examples. A number of weakly-supervised learning approaches have recently been proposed to address classification with limited labelled examples. These weakly-supervised learning approaches can be used to adjust or *correct* classifiers in order to improve their performances when in the presence of many unlabelled instances. Apart from the PU learning approach introduced earlier, there exist many other techniques that address the classification task with weakly supervised examples. In the following, we summarise and discuss such approaches using the categorisation of [19]: (1) **Semi-supervised classification** [5], which focuses on leveraging a small amount of labelled examples to improve the performance of unsupervised learning and requires reliable examples for both positive and negative classes; (2) **One-class classification** [21], which focuses on distinguishing the properties of the selected class versus other classes in multi-class classification scenarios and is mainly applied to anomaly detection [19]; (3) **Positive-unlabelled classification** [14], which is frequently adopted to address the problem of insufficient labelled examples of one class in binary classification. This approach is particularly related to our scenario consisting of binary classification with limited positive instances and unlabelled data. However, most of the positive-unlabelled classification approaches require an extra class prior estimation of the positive and unlabelled classes, which leads to a systematic estimation error [12]; (4) **Label-proportion classification** [33], which conducts classification according to the known class distribution. However, the class distribution is missing in our

[1]The definition of sufficiency relies on the corresponding helpfulness threshold.

scenario; (5) **Unlabelled-unlabelled classification** [13], which is akin to clustering and ignores the information of the labelled examples; (6) **Complementary-label classification** [18] leverages extra attributes or information that denote the unrelated pattern of the corresponding class to help conduct multi-class classification; (7) **Similar-unlabelled classification** [2] relies on the pairwise similarity of examples in one class. However, the classification will be inaccurate if the instances were wrongly labelled; (8) **Positive-confidence classification** [19] conducts binary classification with only positive examples and trains the classifier according to positive examples with different levels of confidence.

Our approach in this paper is similar to the positive-unlabelled and positive-confidence classification approaches. However, unlike the positive-unlabelled classification approach, our NCWS approach does not require a class prior estimation. To illustrate the importance of this difference, we use a positive-unlabelled learning-based approach (i.e. C-PU [12]) as a baseline to compare with our NCWS approach. In addition, unlike the positive-confidence classification approach, we instead train the classifier by leveraging the probability of the unlabelled instances to be negative. For evaluation purposes, we use the positive-confidence classification approach (i.e. P-conf [19]) as another baseline to our NCWS approach.

Note that apart from the weakly-supervised learning, Veropoulos et al. [40] also proposed one approach that adjusts the SVM hyperplane by putting a higher penalty on misclassifying the positive class than the negative class. However, this leads to over-sensitivity towards the positive examples when applying the bias penalty strategy as well as an improper boundary shape especially with sparse positive examples [1]. Moreover, we argue that over-relying on the positive examples and ignoring information behind the negative class can lead to an inaccurate classification. Even though the penalty-based approach has these aforementioned disadvantages when doing binary classification, we still use it as another baseline in our experiments and refer to it as 'SVM-P'. As introduced in Section 1, we use the review helpfulness classification as a test case to examine the effectiveness of NCWS. Therefore, we now describe the recent literature on review helpfulness classification.

2.2 Review Helpfulness Classification

In the literature, various helpfulness classification studies have investigated different representations and properties of reviews. We classify such prior studies into five categories with different feature types, including structural features, lexical features, syntactic features, metadata features, and contextual features.

Structural Features: These features capture the structure and formatting of user comments. Many studies consider structural features as strong features in detecting helpful reviews. Liu et al. [26] and Lu et al. [29] leveraged the average sentence length and the number of sentences; Kim et al. [22] explored the effectiveness of the length of comments, the percentage of question sentences and the number of exclamation marks. In general, these studies agree that the length of reviews is one of the most effective features in review helpfulness classification.

Lexical Features: Lexical features analyse the words used in the comments. Kim et al. [22] and Tsur and Rappoport [39] used the TF-IDF score of each word and each bigram as features. Moreover, many deep learning approaches use word embedding for word representations. Word embedding-based approaches have been adopted in [6, 8] and have been shown to have a better expressiveness than other hand-crafted features.

Syntactic Features: These features capture the linguistic properties of user comments. Kim et al. [22] investigated the effectiveness of different syntactic tokens including the percentage of nouns, the percentage of verbs, the percentage of adjectives and the percentage of adverbs. In addition, sentiment words were considered by Yang et al. [44], who obtained significant improvements compared to using simple lexical features.

Metadata Features: These features focus on the relationship between review helpfulness and user ratings. Both Kim et al. [22] and Huang et al. [17] found a positive correlation between review helpfulness rating and review star ratings.

Contextual Features: These features mainly focus on the behaviour of review writers as well as the connection between the review writer and readers. Huang et al. [17] investigated the historical review helpfulness ratings of the review writers and Lu et al. [29] studied the influence of the connection between the review writers and readers on the review helpfulness ratings. They concluded that the contextual features contribute to enhance the accuracy of the review helpfulness classification. However, these contextual features are harder to obtain than the other previous features.

In this paper, we consider some representative features from the aforementioned listed feature sets to develop several review helpfulness classifiers, which we use as basic baselines. Moreover, we include two neural network-based classifiers as additional basic baselines. These basic baselines are then corrected to alleviate the incompleteness of positive labels and the many unlabelled instances using our proposed NCWS approach in comparison to the current weakly supervised SVM-P, C-PU and P-conf approaches from the literature. In the following section, we introduce a comprehensive description of our NCWS approach.

3 METHODOLOGY

First, we introduce the problem of binary classification with limited positive and abundant unlabelled instances, as well as the used notations. Then, we illustrate our proposed NCWS approach and how we derive the loss functions for the positive and negative classes, respectively.

3.1 Problem Statement

The binary classification task consists in identifying the positive instances in a corpus of instances using binary classification approaches. This problem consists of two main objects, the set of instances $X = \{x_1, x_2, ..., x_N\}$ of size N and its corresponding class label set $Y = \{y_1, y_2, ..., y_N\}$ with label $y_i \in \{+1, -1\}$. Our objective is to obtain an unbiased and accurate classifier $g(x) \rightarrow \{+1, -1\}$, which can accurately identify the positive and negative instances by modelling the limited positive and abundant unlabelled instances.

In our scenario, a -1 label indicates that the instance is unlabelled, rather than being negative. For example, in the review helpfulness classification test case, unlabelled reviews could be the result of a number of reasons, such as when the review is not yet old enough to have gained sufficient viewers to provide it with helpful votes, or when the user interface may have not yet shown the review to users [28]. For this reason, review helpfulness classification can be seen as an example of classification with limited positive instances

and many unlabelled instances. Next, we define our NCWS approach, which leverages the properties of the unlabelled instances during classification. We use the number of days d since the review has been posted (i.e. its age in days) in addition to the content of the review to infer confidence estimates about the unlabelled reviews being actually negative. We validate the reliability of using the age of reviews as an adequate instance property in Section 5.2.

3.2 Negative Confidence-aware Weakly Supervised approach (NCWS)

As introduced in Section 1, weak supervision provides more data to the learner. In this paper, we apply weak supervision to a classifier to address the limited positive instances and the preponderance of unlabelled instances. For example, in the review helpfulness classification task, we might reasonably assume that some unlabelled newer reviews may in fact be positive, but have not yet experienced sufficient exposure to users to gain helpful votes; conversely older unlabelled reviews are less likely to be helpful.

To address the likelihood of unlabelled instances belonging to the positive class, we propose a notion of *negativity*, the likelihood that an unlabelled instance belongs to the (latent) negative class. In review helpfulness classification, we assume that the likely negativity of an unlabelled review increases with its age (i.e. the time the review has been posted). This is motivated by the assumption that an old review has a higher probability to receive helpful votes [25, 28, 42]. Therefore, the longer time that an unlabelled review has been posted, the more likely that the review will be unhelpful[2]. Indeed, negativity is orthogonal to the notion of positivity – used by the PU learning approach of Ishida et al. [19] that only uses positive examples – which is the confidence in the positive examples actually belonging to the positive class. In contrast, our approach models the *unlabelled* instances – i.e. *NCWS* – based on properties of these instances that indicate the confidence that they are indeed negative instances. As mentioned in Section 3.1, in the review helpfulness classification case, we use the age of reviews in addition to their content. In the following, we use age to describe and illustrate how we generate the negativity scores during classification.

Firstly, let the negativity score $n(x) = p(y = -1|x)$ for each unlabelled review x be a function of the number of days since the review has been posted ($d(x)$):

$$n(x) = \frac{log(d(x)+1)}{log(max(d(X))+2)} \tag{1}$$

where $max(d(X))$ indicates the age in days of the oldest review in the set of review instances X. Indeed, we argue that the longer that an unlabelled review has been posted, the more likely that the review will be unhelpful. We normalise the value of review post days $d(x)$ into the range $(0, 1)$. A higher negativity score for a review denotes a larger probability that the review is unhelpful for users. Furthermore, let π_+ and π_- indicate the class priors $p(y = +1)$ and $p(y = -1)$, respectively.

Next, our NCWS classification approach builds a classifier, $g(x)$, by minimising the binary classification risk $R(g)$. Let the generic form of a classifier's risk $R(g)$ be as follows:

$$R(g) = E_{p(x,y)}\Big[\ell(y,g(x))\Big] \tag{2}$$

where $E_{p(x,y)}$ indicates the expectation over $p(x, y)$ (i.e. the probability density of instance x for the corresponding label $y \in \{+1, -1\}$), while $\ell(\cdot)$ denotes the loss function of the classifier.

Similar to the usage of *example positivity* in [12], we leverage and incorporate the instance's negativity into the risk function (i.e. Equation (2)) as follows. First, we represent the risk function with the positive and negative prior probabilities as follows:

$$\begin{aligned} R(g) &= E_{p(x,y)}[\ell(yg(x))] = \sum_{y=\pm1}\int \ell(yg(x))p(x|y)p(y)dx \\ &= \int \ell(g(x))p(x|y=+1)p(y=+1)dx \\ &+ \int \ell(-g(x))p(x|y=-1)p(y=-1)dx \\ &= \pi_+E_+\Big[\ell(g(x))\Big] + \pi_-E_-\Big[\ell(-g(x))\Big] \end{aligned} \tag{3}$$

The sum of the posterior probabilities of the two classes can be represented by the negative class-based posterior probability:

$$\begin{aligned} \pi_+p(x|y=+1) + \pi_-p(x|y=-1) &= p(x,y=+1) + p(x,y=-1) \\ = p(x) = \frac{p(x,y=-1)}{p(y=-1|x)} &= \frac{\pi_-p(x|y=-1)}{n(x)} \end{aligned} \tag{4}$$

Therefore, we can represent the positive part with the negative-based probabilities as follows:

$$\begin{aligned} \pi_+p(x|y=+1) &= \frac{\pi_-p(x|y=-1)}{n(x)} - \pi_-p(x|y=-1) \\ &= \pi_-p(x|y=-1)(\frac{1-n(x)}{n(x)}) \end{aligned} \tag{5}$$

According to Equation (5), we can generate the positive summand of the risk function in Equation (3) as follows:

$$\begin{aligned} \pi_+E_+\Big[l(g(x))\Big] &= \int \pi_+p(x|y=+1)\ell(g(x))dx \\ &= \int \pi_-p(x|y=-1)(\frac{1-n(x)}{n(x)})\ell(g(x))dx \\ &= \pi_-E_-\Big[(\frac{1-n(x)}{n(x)})\ell(g(x))\Big] \end{aligned} \tag{6}$$

We then follow the strategy proposed by du Plessis et al. [12] in modelling two classes with distinct loss or risk functions. While for the positive class we retain the original risk function (Equation (2)), for the negative class, we combine Equations (3) and (6) as follows:

$$R(g) = \pi_-E_-\Big[(\frac{1-n(x)}{n(x)})\ell(g(x)) + \ell(-g(x))\Big] \tag{7}$$

Finally, we implement the risk function with the following objective function for the positive and negative instances respectively:

$$R(g) = \begin{cases} min\sum_{i=1}^{n}\Big[\ell(g(x_i))\Big], & \text{if } y_i = 1 \\ min\sum_{i=1}^{n}\Big[(\frac{1-n(x)}{n(x))})\ell(g(x)) + \ell(-g(x))\Big], & \text{otherwise} \end{cases} \tag{8}$$

Thus far, we have formally introduced our NCWS approach with the aforementioned objective function (Equation (8)). Note that this is a general definition and can be applied to various generic binary classification approaches, and moreover, to various classification tasks for which a negativity score $n(x)$ can be defined (such as the age of review for review helpfulness). In this paper, we apply NCWS to the loss function in an SVM classifier as well as two other neural network-based classifiers, specifically classifiers based on

[2]We further validate the underlying assumption in our used datasets in Section 5.2.

Table 1: Categorised hand-engineered features for SVM.

Structural Features	
LEN	The number of words included in each review.
NoS	The number of sentences contained in each review.
ASL	Average sentence length in each review.
PoQS	Percentage of question sentences in each review.
Structural	Combines all features in this structural feature category.
Lexical Feature	
UGR	Unigram, uses TF-IDF to generate a document feature vector for each review.
Syntactic Feature	
Syn	The percentage of nouns, adjectives and adverbs in each review
Metadata Features	
Rating	The review's corresponding rating value
Rating-Norm	Normalised rating - the difference between the average rating of the review's user and the corresponding rating for each review.
Age	The number of days since the review was posted (normalised using Eq. (1)).
ALL	This combines all hand-engineered features into one integrated feature.

CNN and BERT. Next, we introduce the used classifiers for review helpfulness classification.

4 REVIEW HELPFULNESS CLASSIFICATION

In the following experiments, to demonstrate the generalisation of our NCWS approach, we use two families of classifiers. One is based on SVM along a set of hand-engineered features commonly used in the literature. We also use two neural network-based classifiers, namely CNN and BERT-based classifiers.

SVM with Hand-Engineered Features: We use a support vector machine (SVM) to classify users' posted reviews into helpful and unhelpful classes. Similar to Kim et al. [22], we use four groups of features in the SVM-based classifiers, namely Structural, Lexicon, Syntactic, and Metadata features. Table 1 lists the 11 applied features. To further enhance the basic SVM-based classifiers, we add another metadata feature, namely Age to the features listed in Section 2.2. Age is a feature leveraged in our proposed NCWS approach as a key property of a review. Hence, it is added to the basic classifiers to provide additional insights into the performance of NCWS. Finally, we combine all the features together into a single feature (ALL) that comprehensively considers all review's information. For ease of notations, we denote by 'SVM-X', an SVM classifier that uses the list of features X (e.g. 'SVM-LEN' denotes the SVM classifier that is based on the LEN feature while SVM-ALL is the classifier that that considers all features).

Neural Network (NN) Classifiers: We use CNN and BERT-based classifiers to experiment with the performance of state-of-the-art classifiers using neural-network-based approaches to validate the effectiveness of NCWS on different classifier types. We describe the implementation details of these two classifiers in Section 5.3.

5 EXPERIMENTAL SETUP

In this section, we formulate our research questions, depict the datasets we use and provide details on the experimental setup of the baselines, our NCWS approach as well as the evaluation metrics.

5.1 Research Questions

In the following, we evaluate the usefulness of our NCWS approach in comparison to other weakly supervised or other classification correction approaches from the literature (the *baselines*). First, we validate the effectiveness of the review helpfulness classification approaches introduced in Section 4, which are our *basic* classifiers:

RQ1: How effective are the review helpfulness classification approaches? (Section 6.1)

Table 2: Summary of the used datasets.

Dataset	#Reviews	#Helpful	#Unlabelled	%Helpful
Yelp	1,373,587	621,112	770,780	45.21%
Kindle	982,619	407,019	575,600	41.42%
Electronics	1,689,188	633,154	1,056,034	37.48%

Second, we evaluate our proposed NCWS approach applied to those selected basic classifiers:

RQ2: Can our proposed NCWS approach outperform other classification correction approaches on all used datasets (i.e. Yelp and Amazon) and can it generalise to classifiers with differing performances? (Section 6.2)

Finally, we demonstrate the usefulness of the review helpfulness models attained using NCWS to the application of venue recommendation, as per our final research question:

RQ3: Does an improved NCWS-based review helpfulness classifier benefit an existing state-of-the-art review-based venue recommendation model? (Section 7)

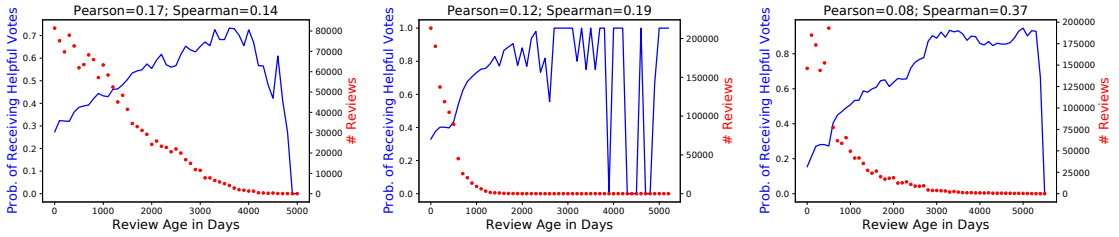
5.2 Datasets

To address RQ1 & RQ2, we conduct experiments on three datasets from two data sources, namely Yelp, and the Amazon. We use the top categories from each data source with the largest amount of user reviews. Such a filtering strategy alleviates data sparseness in those categories with few user reviews, focusing instead on categories with a rich user feedback. For Yelp, we use the Yelp dataset challenge round 12[3]. After that, we collect reviews from the top three categories including 'restaurants', 'food', and 'nightlife'. For Amazon, we use a popular public Amazon review dataset [15], which has been adopted in prior review helpfulness studies [11, 30]. We select reviews from two popular categories, namely 'Kindle' and 'Electronics' as our two Amazon datasets.

In the Amazon-based datasets, the user information is missing, hence for these datasets, only the Rating feature is present in the Metadata feature class. As a consequence, we examine the performances of 13 and 12 classifiers on the Yelp and Amazon datasets, respectively. Specifically, recall that we refer to the classification approaches we introduced in Section 4 as the basic classifiers. For these three datasets, following [24], we set the review helpfulness threshold to 1, i.e. we consider reviews to belong to the positive class only if they have one or more helpful votes, otherwise, the reviews are regarded as unlabelled. Table 2 provides statistics for these datasets.

We conduct a 5-fold cross-validation on the three used datasets. It is apparent from the datasets summary in Table 2 that these datasets are imbalanced, having all a smaller number of helpful reviews than that of the unlabelled reviews. Therefore, we balance the class distribution of the training dataset by applying a down-sampling strategy. Down-sampling is a well-known solution to address the class distribution imbalance when training binary classifiers [37]. In particular, as introduced in Section 3, we take the age of reviews into account during the calculation of the review negativity by leveraging the assumption that an older review has a higher probability to have received helpful votes. To validate our assumption on the used datasets, in Figures 1(a) - 1(c), we plot the age of reviews (in days) against the probability that reviews of that age have been labelled as helpful. In the same figures, we also plot the review frequency by age (in red).

[3] https://www.yelp.com/dataset/challenge



(a) Yelp **(b) Kindle** **(c) Electronics**

Figure 1: The probability of obtaining helpful votes for reviews with different number of days (age) since a review was posted (blue lines) and the number of posted reviews of different ages (red dots), for the Yelp, Kindle and Electronics datasets.

From Figures 1(a) - 1(c), we observe that the Yelp, Kindle and Electronics datasets share similar helpful vote distributions across review ages. Indeed, on all three datasets, these plots appear to corroborate our assumption that there is a correlation between the number of received helpful votes by a review and the number of days the review has been posted, since the older reviews have a higher probability of receiving helpful votes than the younger reviews. In particular, we calculate the Pearson and Spearman correlations between the age of reviews (in days) and the helpful vote probability. According to the value of correlation scores, all three datasets exhibit positive Pearson and Spearman correlation scores[4]. This statistically validates our assumption that older reviews have a higher probability of receiving helpful votes. Note that for the Kindle dataset, reviews that are more than 2000 days old are rare (see the corresponding frequency plot using red dots), explaining the high variance in the corresponding helpful vote probabilities.

Armed with these (weak) correlations, the underlying assumption of NCWS is thus: the longer a review has been posted that remains unlabelled, the more likely that the review will be unhelpful. Indeed, new reviews without votes could be helpful, but have not had sufficient opportunity to be presented to users; on the other hand, older reviews have not received helpful votes despite being presented to users. Therefore, to evaluate the effectiveness of NCWS and its underlying assumption, we apply NCWS to two generic types of classifiers, namely the SVM models with the features introduced in Section 4 and the NN-based classifiers. As mentioned in Section 4, we also use the age of reviews as an additional feature for the SVM classifier (denoted SVM-Age). Because of the observed correlations between the helpfulness and the age of reviews and the reliance of NCWS on the age property, such a basic baseline allows to evaluate if any improvement is the result of the use of the age feature itself, or whether it is due to NCWS itself. We also compare the resulting classification performances to the same classifiers but corrected using the competing methods from the literature, namely SVM-P, C-PU, and P-conf. We describe these classifiers and their corrected versions in the next section.

5.3 Classifiers

We use the three following so-called basic classifiers:

SVM: We implement the SVM model with the LIBSVM [4] library. Moreover, we use the default setting for the parameter values with a penalty parameter C = 1.0 and the RBF kernel. We instantiate different SVM classifiers based on the features sets listed in Table 1.

[4]All correlations are statistically significant according to the scipy implementations.

CNN: The CNN-based classifier consists of three vertically concatenated convolutional layers. Each layer has different convolutional filter sizes (namely $3 \times m$, $4 \times m$ and $5 \times m$ respectively, where m is the embedding size of a word). The output of the third convolutional layer is then fed into a linear prediction layer to predict the review helpfulness label. Moreover, we use the public pre-trained word2vec vectors from Glove [32] with $m = 100$. In particular, we adopt the cross-entropy loss to train the CNN model.

BERT: We implement a BERT-based classifier with a popular natural language processing architecture (from HuggingFace's Transformer [43] library), which enables a quick implementation of the BERT transformer to process text. In particular, we adopt the pre-trained BERT model (i.e. 'bert-base-uncased'). Next, following the setup of the CNN-based classifier, we again use a linear prediction layer to make the review helpfulness predictions and train the model by using the classic cross-entropy loss function. Both NN-based classifiers are trained using batch size 100 for 10 epochs, using the Adam optimiser with a learning rate of 10^{-4}.

Classifier Correction Baselines: Our experiments also apply three existing correction approaches (namely SVM-P, C-PU and P-conf) as baselines to correct the basic classifiers:

• **SVM-P:** The SVM-P approach applies a larger penalty value to the positive class than to the negative class according to the ratio of the number of positive examples versus the negative examples as suggested by Tang et al. [38]. We use LIBSVM [4] to apply the penalty values to different classes in the loss function. However, the experimental setup of SVM-P is different from other classification approaches in its training process. Indeed, the penalty values for different classes correspond to the class ratios of an unbalanced dataset. Therefore, we calculate the class ratio of the training dataset for SVM-P before down-sampling. In particular, the SVM-P approach is limited to the SVM-based classifiers.

• **C-PU:** This approach was proposed by du Plessis [12]. As mentioned in Section 2, C-PU has a similar methodology to our NCWS approach, applying different loss functions for the positive and unlabelled examples. However, unlike our approach, C-PU does not consider the negative confidence of unlabelled instances. For SVM, following [12], we use the double hinge loss, $\ell_{dh}(z) = max(-z, max(0, \frac{1}{2} - \frac{1}{2}z))$, when applying C-PU. For the NN classifiers, we directly integrate C-PU into the cross-entropy loss function.

• **P-conf [19]:** As introduced in Section 2, P-conf learns a classifier only from the positively-labelled instances and leverages the probability of these instances to be positive. We apply the objective function of P-conf [19] to all basic classifiers.

5.4 Evaluation Metrics

In this paper, we aim to detect positive examples with weakly supervised binary classifiers in the review helpfulness classification task. Therefore, we use the F1 metric as the key metric to evaluate the performances of the classifiers in accurately classifying the reviews. Precision and recall are also reported to further examine the classification accuracy and the models' ability to identify positive examples in the corpus.

It is of note that a number of studies have proposed approaches for the evaluation of PU learning. Claesen et al. [9] proposed to use a ranking-based evaluation approach and set the threshold value to divide the positive and negative examples. Jain et al. [20] proposed to evaluate the performance of PU learning-based classifiers with the aid of the class prior knowledge of the class distribution in the unlabelled dataset. However, the evaluation approaches of [9, 20] require the estimation of information such as the class threshold and the class prior, which causes a systematic estimation bias in the evaluation process [12]. Therefore, we resort to using the classical evaluation metrics we introduced above and rely only on the ground truth of the positive examples that we have.

6 RESULTS ANALYSIS

In this section, we present and analyse the results of our experiments and answer the first two research questions in Section 5.1. These research questions focus on identifying the review helpfulness classifiers with the best performances among the basic classifiers and enable with higher reliability to examine the effectiveness of our proposed NCWS approach along several classifiers of differing performances in comparison with other classification correction approaches, namely SVM-P, C-PU and P-conf.

6.1 RQ1: Review Helpfulness Evaluation

To address RQ1 and identify the best performing basic classifiers, we compare the effectiveness of various basic classifiers in distinguishing between helpful and unhelpful reviews using the F1 score. The results over the three used datasets are presented in Table 3.

From Table 3 we observe that, among the SVM-based classifiers, the SVM-LEN, SVM-Structural and SVM-ALL classifiers outperform other classification approaches and provide the best classification performances. They obtain the highest F1 scores across the 3 datasets (>0.6 on the Yelp and Electronics datasets and >0.45 on the Kindle dataset). Meanwhile, SVM-NoS and SVM-Age also obtain good classification performances on the Yelp and Electronics datasets (F1 scores >0.5). However, their classification performances decrease on the Kindle dataset (<0.35). In particular, by observing the good performances of the classifiers that deploy review length as a feature (i.e. SVM-LEN, SVM-NoS, SVM-Structure and SVM-ALL), we conclude that the length of a review is a useful feature for predicting review helpfulness. Furthermore, the unstable performances of the SVM-Age classifier indicates that the age feature cannot by itself fully address the review helpfulness classification problem by leveraging the (weak) correlations between the age and the helpfulness of review that was validated in Section 5.2. Apart from these discussed classifiers, the remainder of the SVM-based classifiers each obtains high F1 scores on some but not all of the 3 datasets or exhibits bad performances across the 3 datasets. For example, the SVM-Rating classifier obtains good classification results on the Yelp and Electronics datasets but obtains a very low F1 score on the Kindle dataset. On the other hand, the NN-based classifiers also provide good results with high F1 scores (>0.48). In particular, the BERT classifier obtains higher F1 scores than the CNN classifier across all 3 datasets. However, the best performing SVM-based approach (i.e. SVM-ALL) still outperforms these two NN-based approaches on all the 3 datasets.

Table 3: Performances of the basic classifiers.

Basic Classifiers	Yelp	Kindle	Electronics
SVM-LEN	0.6069	0.4679	0.6047
SVM-NoS	0.5975	0.3362	0.5759
SVM-ASL	0.4931	0.0432	0.5315
SVM-PoQS	0.2153	0.0132	0.5287
SVM-Structural	0.6017	0.4576	0.6013
SVM-UGR	0.5344	0.0354	0.0925
SVM-Syn	0.0846	0.0086	0.0215
SVM-Rating	0.5439	0.0753	0.5081
SVM-Rating-Norm	0.5843	-	-
SVM-Age	0.5428	0.3037	0.5451
SVM-ALL	**0.6340**	**0.5877**	**0.6336**
CNN	0.5018	0.4830	0.5103
BERT	0.6119	0.5618	0.5712

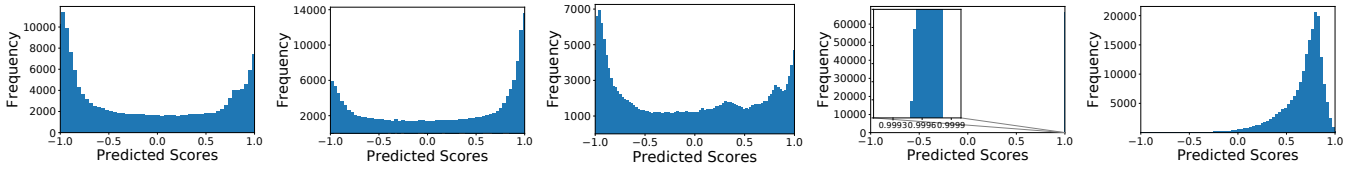
Therefore, for RQ 1, we conclude that the basic classifiers that account for the length of a review, including the SVM-LEN, SVM-NoS, SVM-Structural and SVM-ALL classifiers, provide the best overall performances among the SVM-based classifiers. This conclusion highlights the effectiveness of taking review length into account in the review helpfulness classification task. This is in line with their good performances reported in the literature - see Section 2.2. Moreover, the age feature-based classifier (i.e. SVM-Age) also shows reasonable classification results, which is in line with our observed correlations between the age and the helpfulness of reviews. Furthermore, the NN-based approaches (i.e. the CNN and BERT-based classifiers) lead also to competitive review helpfulness classifiers. As a consequence, we use these aforementioned SVM and NN-based classifiers as representatives of reasonable review helpfulness classification approaches and for evaluating the effectiveness of our proposed NCWS approach.

6.2 RQ2: Classification Correction Evaluation

For RQ 2, we examine the effectiveness of our NCWS approach along the selected basic classifiers from the previous section's conclusions and in comparison to other existing classification correction approaches (namely SVM-P, C-PU and P-conf) from the literature, on three datasets, namely, the Yelp, Kindle and Electronics datasets. As mentioned in Section 1, we aim to address the problem of binary classification with incomplete positive instances and abundant unlabelled instances. Following the general weak supervision paradigm, the main objective of our NCWS approach is to help classifiers model the unlabelled instances by identifying further positive instances[5] from the many unlabelled instances, thereby improving their performances.

First, we compare the differences between the classification results of the basic classifiers and the results after applying the corresponding classification correction approaches to the basic classifiers. Figures 2 plots the frequency distributions of the predicted scores of review instances (in the range -1 to 1), to illustrate the alteration of the classification results when NCWS and the classification correction baseline approaches are deployed. For reasons of brevity, we show the classification results of the SVM-ALL classifier, the best performing classifier, as a representative example of the effects of applying NCWS and the various classification correction methods

[5]These instances would have otherwise been assumed to be negative.



(a) SVM-ALL basic classifier (b) NCWS approach (c) SVM-P approach (d) C-PU approach (e) P-conf approach

Figure 2: Frequency distribution of review helpfulness score predictions for the SVM-ALL basic classifier and the corresponding classification correction approaches applied to the SVM-ALL basic classifier.

Table 4: Results of the classification correction approaches on the Yelp, Kindle and Electronics datasets. Statistically significant differences, according to the McNemar's test ($p < 0.05$), to the corresponding basic classifier are indicated by *.

		Yelp				Kindle				Electronics			
		Precision	**Recall**	**F1**		**Precision**	**Recall**	**F1**		**Precision**	**Recall**	**F1**	
SVM-LEN	basic	0.5781	0.6421	0.6069		0.5369	0.4243	0.4679		0.5203	0.7218	0.6047	
	SVM-P	0.5661	0.6791	0.6169		0.5405	0.3742	0.4288	*	0.5125	0.7373	0.6047	
	NCWS	0.5503	0.7258	**0.6256**	*	0.5294	0.4890	**0.5083**	*	0.5113	0.7421	**0.6054**	*
SVM-NoS	basic	0.5597	0.6437	0.5975		0.5370	0.2603	0.3362		0.5268	0.6351	0.5759	
	SVM-P	0.5418	0.6930	0.6081		0.5377	0.2421	0.3139		0.5268	0.6351	0.5759	
	NCWS	0.5315	0.7330	**0.6151**	*	0.5345	0.2799	**0.3443**	*	0.4777	0.7468	**0.5827**	*
SVM-Structural	basic	0.5819	0.6309	0.6017		0.5364	0.4103	0.4576		0.5386	0.6805	0.6013	
	SVM-P	0.5790	0.6378	0.6036		0.5392	0.3790	0.4343		0.5462	0.6621	0.5986	
	NCWS	0.5501	0.7263	**0.6252**	*	0.5267	0.4883	**0.5011**	*	0.5169	0.7295	**0.6051**	*
SVM-Age	basic	0.5569	0.5293	0.5428		0.6457	0.1986	0.3037		0.6012	0.4987	0.5451	
	SVM-P	0.5925	0.3990	0.4769		0.7338	0.0840	0.1508		0.6598	0.3336	0.4431	
	NCWS	0.5159	0.7098	**0.5975**	*	0.5910	0.2526	**0.3539**	*	0.5939	0.5243	**0.5569**	*
SVM-ALL	basic	0.6023	0.6253	0.6136		0.5254	0.6023	0.5612		0.5975	0.6619	0.6280	
	SVM-P	0.5651	0.6932	0.6226		0.5148	0.6237	0.5641		0.5472	0.6974	0.6132	
	NCWS	0.5751	0.7063	**0.6340**	*	0.4988	0.7152	**0.5877**	*	0.5730	0.7086	**0.6336**	*
CNN	basic	0.5416	0.4674	0.5018		0.4969	0.4699	0.4830		0.4291	0.6294	0.5103	
	SVM-P	-	-	-		-	-	-		-	-	-	
	NCWS	0.5291	0.5254	**0.5272**	*	0.4718	0.5362	**0.5019**	*	0.4082	0.7624	**0.5317**	*
BERT	basic	0.5248	0.7338	0.6119		0.4783	0.6807	0.5618		0.5161	0.6395	0.5712	
	SVM-P	-	-	-		-	-	-		-	-	-	
	NCWS	0.5034	0.7927	**0.6157**	*	0.4543	0.7653	**0.5701**	*	0.4923	0.7127	**0.5823**	*

on the Kindle dataset. However, the observed trends remain consistent across all basic classifiers for the other 2 Yelp and Electronics datasets. In particular, Figure 2(a) shows the review helpfulness score predictions given by the basic SVM-ALL classifier, while Figure 2(b) shows the obtained predictions after correcting the classifier using our NCWS approach. It is clear from the figures that when NCWS is applied, a further number of reviews are classified as positive instances by the corrected SVM-ALL classifier. These results are in line with the objective of our NCWS approach to identify more positive examples from the unlabelled instances. Figures 2(c)-(e) show that after applying the classification correction baseline approaches (i.e. SVM-P, C-PU and P-conf) to the SVM-ALL classifier, the helpfulness score prediction distributions of the reviews change in different ways. While SVM-P adjusts the classification results of the basic SVM-ALL classifier by identifying further positive review instances, C-PU and P-conf completely change the frequency distribution of the original SVM-ALL classifier's score predictions. In particular, they classify most review instances as positive with C-PU squeezing most of the classification results into an extremely narrow range (i.e. between 0.9994 and 0.9998). As a consequence, only NCWS and SVM-P are useful at correcting and enhancing the performance of the SVM-ALL classifier. Hence, in the following, we focus our experiments on the best performing basic classifiers – as identified in the conclusions of RQ1 – and the corresponding NCWS and SVM-P corrected results, on the 3 datasets, to further examine their overall effectiveness.

Table 4 examines the effectiveness of SVM-P and NCWS after applying them to the SVM and NN-based classifiers on the 3 datasets. For the SVM-P approach, the F1 scores show that SVM-P can be helpful to all the basic classifiers – except SVM-Age – and enhances their classification performance with higher F1 scores. However, such effectiveness is not generalisable to all SVM-based approaches on the 3 datasets. For example, SVM-P negatively impacts the SVM-LEN classifier on the Kindle dataset with a lower F1 score (0.4679 → 0.4288). Meanwhile, our NCWS approach outperforms SVM-P by exhibiting higher F1 scores. In particular, it can significantly and consistently enhance the basic SVM-based classifiers to yield higher F1 scores according to the McNemar's test. On the other hand, apart from these SVM-based classifiers, we further examine the effectiveness of NCWS[6] on the NN-based classifiers, which have been shown to be effective in many neural language processing tasks. Table 4 shows that significant and consistent improvements are observed after applying NCWS to CNN and BERT-based NN classifiers on the 3 datasets. This demonstrates the effectiveness of NCWS when applied to the cross-entropy loss function. Moreover, when applied to the best performing approaches (i.e. SVM-LEN, SVM-Structural, SVM-ALL and BERT), NCWS can further increase the F1 scores of these classifiers. For example, NCWS improves SVM-ALL to obtain the overall best classification performance on the 3 datasets.

Table 5 shows how many unlabelled reviews were classified by the basic classifiers as negative but classified as positive by our NCWS approach. For example, in the Yelp dataset, the SVM-LEN basic classifier predicts 90,260 unlabelled examples as unhelpful while NCWS predicts 7,559 instances of these reviews as being helpful (i.e. an ~8.3% increase). In general, about 5-30% of the unlabelled reviews are re-labelled as positive and identified as helpful by NCWS. These results align with our objective to identify more positive instances from the unlabelled corpus. Therefore, by analysing

[6]The SVM-P method is limited to SVM-based classifiers as introduced in Section 5.3.

Table 5: NCWS's impact on classification: number of unlabelled reviews that changed from being predicted negative to predicted positive by the application of NCWS; total number of reviews predicted negative by the basic classifiers; the percentage of reviews that changed is also shown.

	Yelp		Kindle		Electronics	
SVM-LEN	7559 / 90260	(8.3%)	10030 / 92138	(10.8%)	4587 / 126187	(3.6%)
SVM-NoS	15714 / 97301	(16.1%)	2563 / 104016	(2.4%)	31255 / 138253	(22.6%)
SVM-Structural	20534 / 95746	(21.4%)	6888 / 93687	(7.3%)	12562 / 136656	(9.1%)
SVM-Age	52797 / 160458	(32.9%)	9781 / 171451	(5.7%)	6778 / 232405	(2.9%)
SVM-ALL	23603 / 149574	(15.78%)	20235 / 101828	(19.8%)	21413 / 201927	(10.6%)
CNN	25204 / 171307	(14.7%)	35012 / 119640	(29.3%)	50397 / 152453	(33.1%)
BERT	12503 / 78451	(15.9%)	9217 / 68815	(13.3%)	10512 / 165227	(6.3%)

the results from the three datasets, we can now answer RQ 2: NCWS can successfully improve the performances of basic classifiers while outperforming other classification correction approaches on the F1 evaluation metric. These results validate our assumption that the age of reviews is a reliable signal to infer which of the unlabelled reviews are actually unhelpful and that the longer an unlabelled review has been posted, the more likely this review is unhelpful. In the next section, we show how the performance of a state-of-the-art review-based venue recommendation recommendation system can benefit from the further identified positive instances.

7 VENUE RECOMMENDATION APPLICATION

In this section, we use venue recommendation as a use case to demonstrate the benefit of NCWS, thereby addressing RQ 3 stated in Section 5.1, which focuses on examining the usefulness of the identified helpful reviews in enhancing the review-based recommendation performance. Venue recommendation is an important application of recommendation techniques, where the aim is to suggest relevant venues that a user would like to visit. Often, venue recommendation approaches make use of data from a location-based social network such as Foursquare or Yelp. This data can be implicit feedback, in the form of checkins, or more explicit, such as ratings or reviews [41].

We apply NCWS in the context of a state-of-the-art venue recommendation model, namely DeepCoNN [45], which uses the reviews of users on venues for recommendation as input to a convolutional neural network. DeepCoNN constructs two parallel convolutional neural networks to model user and venue reviews, respectively, to predict the rating that a user would give to a venue. Indeed, DeepCoNN has been frequently used in many review-based recommendation works as a baseline [7, 34].

To integrate NCWS into DeepCoNN, we simply replace the user and venue review corpora with only those reviews that are predicted to be helpful. In doing so, we postulate that removing noisy or unlabelled reviews and replacing them with further positive reviews as identified by NCWS results into a more effective learned DeepCoNN model. We validate this through experiments on the Yelp dataset, which is a widely used venue recommendation dataset [41].

7.1 Experimental Setup

We compare different replacement strategies to assess the effectiveness of the corresponding review selection approaches as input for DeepCoNN. Such review selection approaches include: (1) **+Random:** randomly samples the same numbers of reviews as the predicted helpful reviews with the best corrected classifier (i.e. SVM-ALL) as input for DeepCoNN; (2) **+Basic:** selects helpful reviews with the basic SVM-ALL classifier, which had the best performance in review helpfulness classification in Section 6.1; (3) **+NCWS:** uses the predicted helpful reviews with a NCWS-corrected SVM-ALL classifier. Similarly, as additional comparative approaches, we use the corresponding (4) **+SVM-P**, (5) **+C-PU** and (6) **+P-conf** but with the SVM-P, C-PU and P-conf's predicted helpful reviews instead. We apply these different review-selection strategies within the DeepCoNN venue recommendation model on the Yelp dataset, along with two baselines: DeepCoNN and a popular rating prediction approach, namely Non-negative Matrix Factorisation (NMF) [36], which considers ratings, but not the text of the reviews.

Our experiments are conducted using a 5-fold cross validation, following as closely as possible the experimental setup of [45] (with the same trained word embedding model, but with a different dataset and review selection strategies). We evaluate the rating prediction accuracy using Mean Average Error (MAE) and Root Mean Square Error (RMSE). For both metrics, smaller values are better. We use the paired t-test to determine significant differences of MAE[7].

Table 6: Venue recommendation results: Significant MAE improvements (t-test, $p<0.05$) w.r.t. DeepCoNN, DeepCoNN+Basic & DeepCoNN+SVM-P are denoted by ○, ● & ∗, resp..

	MAE	RMSE
NMF	1.1526	1.4345
DeepCoNN	0.8969	1.1798
+Random	0.9201 ○∗	1.2278
+Basic	0.8629 ○	1.1012
+C-PU	0.8969 ●∗	1.1798
+P-conf	0.8969 ●∗	1.1798
+SVM-P	0.8597 ○	1.0998
+NCWS	**0.8503** ○ ● ∗	**1.0954**

7.2 Results

Table 6 presents the MAE and RMSE scores of the DeepCoNN variants. In particular, the first group of rows are baselines, while the second group corresponds to approaches that make use of review helpfulness classification when filtering the set of reviews to use. On analysing the table, comparing the rating prediction error of DeepCoNN and NMF using the MAE and RMSE metrics, we observe that DeepCoNN, which uses all reviews for venue recommendation enables better representations of user preferences and venue properties with lower rating prediction errors than NMF. Indeed, recall that NMF is only trained on ratings, while DeepCoNN has access to the text of the reviews. In relation to the helpful reviews, we observe that by filtering the reviews to include only those that are predicted to be helpful (c.f. DeepCoNN+Basic, +SVM-P and +NCWS), the rating prediction is improved (i.e. significantly reduced MAE and RMSE scores) compared to DeepCoNN. This implies that, among all the reviews considered by the DeepCoNN baseline, some are noisy, and removing these to focus upon the likely helpful reviews aids learning an effective venue recommendation model. On the other hand, the +C-PU and +P-conf integrations have the same performances as DeepCoNN – indeed, this is expected from the results of Section 6.2, where P-Conf and C-PU were not effective in identifying helpful reviews.

Moreover, comparing the performances between the +Basic, +SVM-P and +NCWS integrations, we find that our proposed NCWS

[7]RMSE, which is a non-linear aggregation of squared absolute errors, is not suitable for significance testing.

approach results in better rating predictions, exhibiting a significant 1.8% improvement in MAE and 0.27% improvement in RMSE, respectively, in comparison to the +Basic approach that uses the basic SVM-ALL classifier. In particular, +NCWS also shows a significantly better performance than +SVM-P, which indicates the benefit of using NCWS over SVM-P in identifying helpful reviews as input for the DeepCoNN model. Moreover, +NCWS is significantly more accurate than +Random, a variant that uses the same number of randomly sampled reviews. Hence, and in answer to RQ3, we find that focusing on the likely helpful reviews, particularly those additional reviews found using our proposed NCWS classifier (see Table 5), allows the performance of the state-of-the-art DeepCoNN rating prediction approach to be significantly enhanced. This also validates the effectiveness of NCWS in identifying helpful reviews.

8 CONCLUSIONS

In this paper, we proposed a novel weak supervised binary classification correction approach by considering the negative confidence of the unlabelled examples under the positive and unlabelled learning scenario. Using three datasets, we showed the effectiveness of our NCWS approach in comparison to several state-of-the-art classification correction approaches from the literature. We also illustrated how NCWS allows to increase the number of positive instances by 5–30% when integrated into various binary classifiers. NCWS is a general classification correction approach, which can be applied to various other classification tasks for which a negativity score can be defined. Using review helpfulness classification as a use case, we extensively demonstrated the effectiveness of NCWS in leveraging the predicted helpful reviews to significantly enhance the performance of DeepCoNN, a recent and strong review-based recommendation model. As future work, we plan to apply NCWS to other binary classification use cases where there are limited labelled examples.

REFERENCES

[1] Rehan Akbani, Stephen Kwek, and Nathalie Japkowicz. 2004. Applying support vector machines to imbalanced datasets. In *Proc. of ECML*.

[2] Han Bao, Gang Niu, and Masashi Sugiyama. 2018. Classification from Pairwise Similarity and Unlabeled Data. In *Proc. of ICML*.

[3] Gilles Blanchard, Gyemin Lee, and Clayton Scott. 2010. Semi-Supervised Novelty Detection. *Journal of Machine Learning Research* 11 (2010), 2973–3009.

[4] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 27:1–27:27.

[5] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien (Eds.). 2006. *Semi-Supervised Learning*. MIT Press.

[6] Cen Chen, Yinfei Yang, Jun Zhou, Xiaolong Li, and Forrest Sheng Bao. 2018. Cross-Domain Review Helpfulness Prediction Based on Convolutional Neural Networks with Auxiliary Domain Discriminators. In *Proc. of ACL*.

[7] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. In *Proc. of WWW*.

[8] Jie Chen, Chunxia Zhang, and Zhendong Niu. 2016. Identifying Helpful Online Reviews with Word Embedding Features. In *Proc. of Knowledge Science, Engineering and Management*.

[9] Marc Claesen, Jesse Davis, Frank De Smet, and Bart De Moor. 2015. Assessing binary classifiers using only positive and unlabeled data. *CoRR* abs/1504.06837 (2015).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR* abs/1810.04805 (2018).

[11] Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and Prediction of Online Product Review Helpfulness: A Survey. In *Proc. of ACL*.

[12] Marthinus Du Plessis, Gang Niu, and Masashi Sugiyama. 2015. Convex formulation for learning from positive and unlabeled data. In *Proc. of ICML*.

[13] Marthinus Christoffel Du Plessis, Gang Niu, and Masashi Sugiyama. 2013. Clustering unclustered data: Unsupervised binary labeling of two datasets having different class balances. In *Proc. of Technologies and Applications of Artificial Intelligence*.

[14] Charles Elkan and Keith Noto. 2008. Learning classifiers from only positive and unlabeled data. In *Proc. of SIGKDD*.

[15] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proc. of WWW*.

[16] Ya-Han Hu and Kuanchin Chen. 2016. Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings. *International Journal of Information Management* 36, 6 (2016), 929–944.

[17] Albert H Huang, Kuanchin Chen, David C Yen, and Trang P Tran. 2015. A study of factors that contribute to online review helpfulness. *Computers in Human Behavior* 48 (2015), 17–27.

[18] Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. 2017. Learning from complementary labels. In *Proc. of NeurIPS*.

[19] Takashi Ishida, Gang Niu, and Masashi Sugiyama. 2018. Binary classification from positive-confidence data. In *Proc. of NeurIPS*.

[20] Shantanu Jain, Martha White, and Predrag Radivojac. 2017. Recovering true classifier performance in positive-unlabeled learning. In *Proc. of AAAI*.

[21] Shehroz S Khan and Michael G Madden. 2009. A survey of recent trends in one class classification. In *Proc. of Artificial Intelligence and Cognitive Science*.

[22] Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *Proc. of EMNLP*.

[23] Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. 2017. Positive-unlabeled learning with non-negative risk estimator. In *Proc. of NeurIPS*.

[24] Srikumar Krishnamoorthy. 2015. Linguistic features for review helpfulness prediction. *Expert Systems with Applications* 42, 7 (2015), 3751–3759.

[25] Pei-Ju Lee, Ya-Han Hu, and Kuan-Ting Lu. 2018. Assessing the helpfulness of online hotel reviews: A classification-based approach. *Telematics and Informatics* 35, 2 (2018), 436–445.

[26] Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. 2007. Low-quality product review detection in opinion summarization. In *Proc. of EMNLP*.

[27] Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. 2008. Modeling and Predicting the Helpfulness of Online Reviews. In *Proc. of ICDM*.

[28] Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. 2008. Modeling and predicting the helpfulness of online reviews. In *Proc. of ICDM*.

[29] Yue Lu, Panayiotis Tsaparas, Alexandros Ntoulas, and Livia Polanyi. 2010. Exploiting social context for review quality prediction. In *Proc. of WWW*.

[30] MSI Malik and Ayyaz Hussain. 2018. An analysis of review content and reviewer variables that contribute to review helpfulness. *Information Processing & Management* 54, 1 (2018), 88–104.

[31] Yoon-Joo Park. 2018. Predicting the Helpfulness of Online Customer Reviews across Different Product Types. *Sustainability* 10, 6 (2018), 1735.

[32] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proc. of EMNLP*.

[33] Novi Quadrianto, Alex J Smola, Tiberio S Caetano, and Quoc V Le. 2009. Estimating labels from label proportions. *Journal of Machine Learning Research* 10 (2009), 2349–2374.

[34] Dimitrios Rafailidis and Fabio Crestani. 2019. Adversarial training for review-based recommendations. In *Proc. of SIGIR*.

[35] Clayton Scott and Gilles Blanchard. 2009. Novelty detection: Unlabeled data definitely help. In *Proc. of AISTATS*.

[36] Suvrit Sra and Inderjit S Dhillon. 2006. Generalized nonnegative matrix approximations with Bregman divergences. In *Proc. of NeurIPS*.

[37] Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. 2009. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence* 23, 04 (2009), 687–719.

[38] Yuchun Tang, Yan-Qing Zhang, Nitesh V Chawla, and Sven Krasser. 2009. SVMs modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, 1 (2009), 281–288.

[39] Oren Tsur and Ari Rappoport. 2009. RevRank: A Fully Unsupervised Algorithm for Selecting the Most Helpful Book Reviews.. In *Proc. of ICWSM*.

[40] Konstantinos Veropoulos, Colin Campbell, Nello Cristianini, and Others. 1999. Controlling the sensitivity of support vector machines. In *Proc. of IJCAI*.

[41] Xi Wang, Iadh Ounis, and Craig Macdonald. 2019. Comparison of Sentiment Analysis and User Ratings in Venue Recommendation. In *Proc. of ECIR*.

[42] Yani Wang, Jun Wang, and Tang Yao. 2019. What makes a helpful online review? A meta-analysis of review characteristics. *Electronic Commerce Research* 19, 2 (2019), 257–284.

[43] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *CoRR* abs/1910.03771 (2019).

[44] Yinfei Yang, Yaowei Yan, Minghui Qiu, and Forrest Bao. 2015. Semantic analysis and helpfulness prediction of text for online product reviews. In *Proc. of ACL*.

[45] Lei Zheng, Vahid Noroozi, and Philip S Yu. 2017. Joint deep modeling of users and items using reviews for recommendation. In *Proc. of WSDM*.

[46] Zhi-Hua Zhou. 2017. A brief introduction to weakly supervised learning. *National Science Review* 5, 1 (2017), 44–53.