Weakly Supervised Learning for Judging the Credibility of Movie Reviews

Han-Sub Shin\textsuperscript{1}, Hyuk-Yoon Kwon\textsuperscript{2} \textasteriskcentered

\textsuperscript{1} Department of Industrial Engineering, Seoul National University of Science and Technology, 232 Gongneung-Ro, Nowon-Gu, Seoul 01811, Korea
\textsuperscript{2} Department of Industrial Engineering and the Research Center for Electrical and Information Technology, Seoul National University of Science and Technology, 232 Gongneung-Ro, Nowon-Gu, Seoul 01811, Korea

Corresponding author: hyukyoon.kwon@seoultech.ac.kr

Abstract

In this paper, we deal with the problem of judging the credibility of movie reviews. The problem is challenging because even experts cannot clearly and efficiently judge the credibility of a movie review and the number of movie reviews is very large. To attack this problem, we propose a weakly supervised learning method for fast annotation. In terms of predefined criteria for weakly supervised learning, we present a simple and clear criterion based on historical movie ratings associated with movie reviewers. The proposed method has the following two advantages. First, it is significantly efficient because we can annotate the entire data sets according to the predefined rule. Indeed, we show that the proposed method can annotate 8,000 movie reviews only in 0.712 seconds. Second, a criterion adapted for weakly supervised learning is simple but effective. We use as a comparison learning method that uses the helpfulness votes of other reviewers as the criterion to judge the credibility of movie reviews, which has been widely used to judge the credibility of online reviews. We indicate that the proposed learning method is comparable to or even better than the helpfulness vote method by showing an improvement over the accuracy of the latter method of 1.57\% \textasciitilde 4.54\%.

Keywords: movie reviews, credibility, machine learning
1 Introduction

A movie review is a subjective evaluation written by a moviegoer, usually consisting of two kinds of data: 1) textual reviews and 2) ratings. Textual reviews are qualitative evaluations; ratings are quantitative evaluations expressed by scores. Fig. 1 shows a sample movie review from IMDb\footnote{https://www.imdb.com/}, a leading website managing movie reviews along with other information. The example shows the rating and the textual review for the movie. We note that these two types of information are closely related: the rating may be considered a numerical summarization reflecting the textual review.

![Sample movie review](image)

Figure 1: Sample movie review

Movie reviews have a significant impact on the decision making of potential moviegoers \footnote{https://www.imdb.com/}. However, a certain group of movie reviewers may generate indistinguishable and/or intentionally malicious reviews so as to bias the overall movie rating. While such groups ought to lack credibility, many well-known existing movie review management systems including Metacritic\footnote{https://www.metacritic.com/} and Rotten Tomatoes\footnote{https://www.rottentomatoes.com} do not consider the credibility of movie reviews in calculating overall movie ratings.

To enhance the evaluation of ratings in movie review systems, it is necessary to make the credibility of movie reviews a factor in the overall movie rating. For this, we need a method to classify movie reviews according to their credibility. However, this is challenging because even experts cannot clearly judge movie review credibility and the number of movie reviews is very large, which requires much time and effort for manual annotation. Table 1 shows an example of different annotation results for a real review from Rotten Tomatoes\footnote{https://www.rottentomatoes.com/}.
Table 1: An example of varying annotation results

| Textual reviews                                                                 | Ratings | Annotation1 | Annotation2 |
|--------------------------------------------------------------------------------|---------|-------------|-------------|
| Really bad. I’m a huge fan of the Marvel movies in general, but this one stands out as a big disappointment. | 4       | Distrusted   | Trusted     |

In this paper, we propose a weakly supervised learning method for judging the credibility of movie reviews. It annotates entire data sets based on a rule-based function instead of manual annotation by human beings, thus annotating at a very fast speed. The most influential factor in the accuracy of weakly supervised learning is a criterion for classification. In this paper, we use a simple and clear criterion based on the historical movie ratings of movie reviewers. Fig. 2 shows actual examples of movie reviewers submitting indistinguishable reviews for multiple movies, which are consequently considered untrustworthy. Fig. 2a is an example where a movie reviewer has given the maximum 10 points to all the movies they have reviewed. We indicate that indistinguishable overall positive reviews with similar expressions are written regardless of the specific movies reviewed. Fig. 2b shows another example where a movie reviewer has given the minimum 0 points and overall negative reviews to all the movies reviewed. Hence, we conclude by evaluating their historical reviews that reviews that will be written by those movie reviewers also have no discrimination.

The contributions of the paper is summarized as follows.

- We propose a weakly learning method for judging the credibility of movie reviews. The proposed method enables fast annotation according to predefined rules. To reduce accuracy limitations in weakly supervised learning, we use a simple but effective criterion to classify movie reviews clearly according to their credibility, which we call **historical credibility**.

- The proposed learning method is **efficient**. Specifically, we show that the proposed method can annotate 8,000 movie reviews in only 0.712 seconds, which occupies about 0.145% of the entire training process when we use Word2Vec as the text representation model. This is a notable result.
Figure 2: Examples of movie reviewers without discrimination

if the annotation results are effective because the usual (human) annotation approach to classifying
movie reviews based on credibility takes much time and effort even for experts.

• The proposed learning model is *effective*. Through extensive experiments, we measure the accuracy
of the proposed learning model and that of the chosen comparison method on 40,000 movie reviews
collected for five different movies. The comparison method uses the helpfulness votes of other
reviewers as the criterion to judge the credibility of movie reviews, a criterion that has been used
in much research. Our finding is that the proposed learning model shows an improvement in
accuracy of the learning model over the comparison method by 1.57% ∼ 4.54% depending on the
text representation model, machine learning technique, and movie genres chosen.

The paper is organized as follows. In Section 2 we describe related work. In Section 3 we describe
text representation models and machine learning techniques that will be used in the proposed learning
method. In Section 4 we propose a weakly supervised learning method using historical credibility as
a criterion for judging the credibility of movie reviews. In Section 5 we explain how the experimental
results show the efficiency and accuracy of the proposed method. In Section 6, we further discuss the proposed learning method. Section 7 concludes the paper.

2 Related Work

Research on weakly supervised learning: Weakly supervised learning allows us to annotate large-scale data at a very fast rate, at the cost of potentially incurring some inaccurate annotation. Typical annotation methods for weakly supervised learning can be classified into three categories: 1) crowdsourcing, 2) distant supervision, and 3) labeling functions. Crowdsourcing is a form of supervised learning, but annotation of data is performed by several annotators who are not experts. Distant supervision extracts the structure from a data set using large-scale existing databases; the extracted structure is used for annotation. Labeling functions are constructed from defined rules for annotating data. This allows annotation to be done quickly, but the results include some noise. The weakly supervised learning used in this paper employs labeling functions because we annotate the entire set of movie reviews based on the predefined criterion, which is obtained from analysis of historical movie ratings.

Recently, weakly supervised learning has been adopted in many studies. For instance, Lee et al. have applied weakly supervised learning to sentiment classification, using weakly supervised learning based on a convolutional neural network to identify keywords for classifying positive and negative sentences. Lin et al. have used weakly supervised learning for understanding the sentiments of users from social media content containing different data models such as texts and images.

Research on movie reviews: For movie reviews, most research has focused on sentiment analysis of textual reviews, because movie reviews depend on the sentiment experienced after watching the movie. Topal et al. have constructed an emotion map according to emotions expressed after watching the movie and have recommended movies to moviegoers based on this emotion map. Manek et al. have performed sentiment classification for movie reviews using a Gini index based on feature selection and a support vector machine (SVM) classifier. Chakraborty et al. have compared the performance of
two clustering algorithms, namely k-means and k-means++, for emotion classification of movie reviews by using Word2Vec as a classifier of the textual review [12]. Alsaqer et al. have improved the accuracy of movie review summarization based on sentiment analysis [13]. He et al. have proposed a self-training method using some labeled features extracted from an existing sentiment lexicon for sentiment analysis [14]. Elmurngi et al. have detected fake movie reviews based on sentiment analysis using supervised learning methods such as SVM and naïve Bayes [15].

**Research on credibility in online review systems:** To the best of our knowledge, there are no public methods that consider the credibility of movie reviews. Many well-known existing movie scoring systems including Metacritic and Rotten Tomatoes do not consider movie review credibility in calculating the overall movie rating. IMDb is known to consider movie reviewers’ credibility; however, since the specific methods are not disclosed, we cannot evaluate their effectiveness.

On the other hand, much effort has been invested in research on review credibility in other online review systems. It is well known that judging the credibility of online reviews is important because they affect consumers’ decision making [16, 17, 18]. Liu et al. have evaluated the effectiveness of reviews of travel products based on two aspects [19]: 1) reviewer information such as profiles and 2) reviews including some quantitative information (i.e., ratings and length of reviews). As a result, they have shown that qualitative judgments are the most influential factors for review credibility. Ghose et al. have explored multiple aspects of textual reviews, namely subjectivity levels, readability, and spelling errors, to understand the helpfulness of reviews in terms of economic and social outcomes [20]. Hochmeister et al. have compared destination experts (i.e., the most active members) in TripAdvisor with general reviewers and have shown that destination experts have more influence in the TripAdvisor community [21]. Barbado et al. have proposed a method for detecting fake online reviews in the domain of consumer electronics businesses (i.e., studying Yelp businesses) [22], labeling reviews in the data set used as trustworthy or fake. Reyes et al. have studied five chosen factors in online reviews that can affect the decision making of customers by exploring TripAdvisor reviews [23].

Some research results have shown that helpfulness votes from other reviewers are a critical measure for evaluating review credibility. Fang et al. have conducted empirical analysis to explore two factors
affecting the value of reviews: 1) the reviews themselves and 2) reviewer characteristics [24], using
the helpfulness vote as the criterion for determining review value. Shan et al. have shown that the
reviewer reputation generated by peer ratings has great influence on credibility evaluation [25]. Yang et
al. have examined six factors (specifically, reviewer location, reviewer level, reviewer helpfulness vote,
review rating, review length, and review photo) and have shown that the review rating and reviewer
helpfulness vote are the most influential factors [26]. Gang et al. have examined the relationship between
emotions and review helpfulness [27]. For this, they have defined three kinds of emotion embedded in
online reviews (specifically, anger, fear, and sadness) and have analyzed the different effects of each
emotion on perceived review helpfulness.

3 Preliminaries

3.1 Text representation models

In this paper, we build a learning model to classify movie reviews based on their credibility. To annotate
movie reviews based on textual content, we need text representation models. We use two kinds of bag-
of-words models: 1) TF-IDF and 2) Word2Vec. TF-IDF is a representative model for representing the
text [28]. Word2Vec is a vector space model that embeds a corpus of text into a vector space [29].
The purpose of this paper is not to propose a specific novel text representation model. Instead, we
choose some representative text representation models to show the effectiveness of the proposed learning
method; we could equally apply any other models in place of TF-IDF and Word2Vec to the proposed
method.

3.1.1 TF-IDF

TF-IDF represents the relative importance of words in a document based on the entire set of documents [30]. Eq. [1]:

is the formula for obtaining TF-IDF, which is calculated by multiplying TF (text frequency) and IDF (inverse document frequency). Essentially, the result of TF-IDF increases as the
frequency of given words in a document increases and the number of documents containing given words decreases. Eq. (2) and Eq. (3) are formulas for obtaining TF and IDF, respectively. TF is the frequency of a particular word in the document. Specifically, \( f(i,j) \) is the frequency of a word \( i \) in a document \( j \) while \( \max \{ f(w,d) : w \in d \} \) is the maximum of \( f(i,j) \) of the words in the document. IDF is the total number of documents \(|D|\) divided by the number of documents containing a word \( t \). IDF uses a logarithm to prevent the result from increasing too much as \(|D|\) increases [31].

\[
TF_{i,j} - IDF_{i,j} = TF_{i,j} \times IDF_i
\] (1)

\[
TF_{i,j} = 0.5 + \frac{0.5 \times f(i,j)}{\max \{ f(w,d) : w \in d \}}
\] (2)

\[
IDF_i = \log \left( \frac{|D|}{\text{count} \{d : t \in d\}} \right)
\] (3)

### 3.1.2 Word2Vec

Word2Vec is a word embedding model that maps words contained in the document into a vector space, where it tries to preserve contextual meanings [29]. Word2Vec supports two kinds of architectures [32]: 1) continuous bag-of-words (simply, CBOW) and 2) continuous skip-gram, as shown in Fig. 3. The continuous CBOW architecture tries to predict the center word from the context words. For example, in a given sentence, “This movie is wonderful,” this architecture tries to find the most likely word to fill in the blank from the context. As a result, it shows better performance when we predict frequently used words. Conversely, the continuous skip-gram architecture tries to predict the context from a given center word. For example, for the given center word “soundtrack,” this architecture tries to predict the surrounding context.
3.2 Machine learning techniques

For the classification of movie reviews, we apply machine learning techniques to the textual reviews. A variety of text classifiers using machine learning techniques have been proposed: decision trees, naïve Bayes, rule induction, neural networks, nearest neighbors, and support vector machines (SVMs).

In this paper, we apply two techniques that have been studied for movie reviews: 1) naïve Bayes and 2) SVM. (As previously noted, we could apply any other machine learning techniques instead of naïve Bayes and SVM for the proposed method.)

3.2.1 Naïve Bayes classifier

The naïve Bayes classifier has been widely used for document classification because it is quite simple but shows high accuracy. The name “naïve Bayes” refers to a classifier based on Bayes’ theorem. The important characteristic of the naïve Bayes classifier is that the features used for the classifier are stochastically independent of each other.

When a set of classes $C = \{c_1, c_2, \ldots, c_n\}$ is given, the naïve Bayes classifier predicts the $c_i$ that is most likely to include a given document $d$ based on Bayes’ theorem. Eq. (4) shows Bayes’ theorem, which gives the probability that a given document $d$ is included in a category $c_i$. $P(d)$ is the probability
of choosing a document $d$ from all the documents. $P(c_i)$ is the number of documents in a category $c_i$ divided by the total number of documents. $P(d \mid c_i)$ is, for a given category $c_i$, the probability of choosing document $d$ in that category. When a set of classes $\{c_1, c_2\}$ is given, we can predict the category in which a document $d$ is included by comparing $P(c_1 \mid d)$ and $P(c_2 \mid d)$.

$$P(c_i \mid d) = \frac{P(c_i, d)}{P(d)} = \frac{P(c_i)P(d \mid c_i)}{P(d)}$$

(4)

### 3.2.2 Support vector machines

A support vector machine (SVM) finds the optimal hyperplane for a linear classification on data on a multidimensional space [36]. Here, we can have multiple possible hyperplanes that can classify the data. The SVM finds a hyperplane having the maximal margin. Fig. 4 shows an example of SVM-based classification. Here, when we compare candidate hyperplanes $B_1$ and $B_2$ in terms of the margin between the classified categories, $B_1$ is chosen as the result hyperplane because its margin is greater than that of $B_2$. The maximal margins concept aims to minimize incorrect classifications.

![Figure 4: An example of the SVM classifier](image)

We can extend SVM to the case of nonlinear classification. For this, the kernel function has
been introduced \cite{37}, defining additional functions for describing new dimensions required in nonlinear classification. Fig. 5 shows the effect of nonlinear classification using the SVM kernel function. Fig. 5a shows objects that are not classified by linear classification in SVM; Fig. 5b shows that they can be classified by nonlinear classification using a kernel function. Kernel functions commonly used for the SVM classifier are the polynomial kernel, the Gaussian kernel, and the radial basis function \cite{37}.

![Figure 5: The effect of nonlinear classification using a kernel function in SVM](image)

(a) Linear classification  
(b) Nonlinear classification

4 The Proposed Method

In this section, we propose a new learning method for judging the credibility of movie reviews. Our research objectives are summarized as follows: the proposed learning model 1) should be efficient at dealing with large-scale movie review data sets and 2) should provide accurate results based on a clear criterion for judging the credibility of movie reviews. To satisfy the first objective, we propose a method based on weakly supervised learning, which allows fast annotation by a predefined rule. To satisfy the second objective, we propose a new criterion for judging the credibility of movie reviews, based essentially on historical rating records. In Section 4.1 we will present the overall framework based on weakly supervised learning. In Section 4.2 we will formally define a new criterion for judging credibility.
4.1 The overall framework

Fig. 6 shows the overall framework for classifying movie reviews based on credibility. It consists of 1) a learning phase and 2) an evaluation phase. In the learning phase, we first analyze historical review ratings for each reviewer so as to determine movie review credibility. In Section 4.2, we present a new criterion for determining movie review credibility. Next, we annotate the entire movie review as “trusted” or “distrusted” based on the proposed criterion. Here, we note that the annotation of the entire data set is performed at a very fast speed because we can annotate movie reviews based on the defined rule using a clear criterion without manual effort. Finally, we build a learning model by obtaining a vector of keywords from textual reviews using two text representation models, namely Word2Vec and TF-IDF, and by applying two machine learning techniques, namely naïve Bayes classifier and SVM, to the annotated data sets. In the evaluation phase, we can use the learning model to judge the credibility of new movie reviews based on the textual review information. Here, thanks to the learning model, we can even judge the credibility of movie reviews that are written by first-time reviewers.

4.2 Historical credibility for weakly supervised learning

In this section we present a simple criterion, which we call historical credibility, for weakly supervised learning. We need a criterion that can clearly classify movie reviews according to their credibility, utilizing the historical movie ratings of each movie reviewer. The basic idea is to identify movie reviewers who have given indistinguishable ratings and to define all movie reviews written by them as distrusted reviews. We define historical credibility as in Definition 1.

**Definition 1.** Historical credibility is a criterion that evaluates review credibility based on the historical review ratings for each reviewer, defined as follows: all reviews are classified as either trusted or distrusted. The distrusted reviews are defined as those written by a reviewer $R$ satisfying the following two conditions: 1) the standard deviation of all past movie ratings $X_R$ is 0 and 2) the average over $X_R$ is either the lowest score (e.g., 1 point) or the highest score (e.g., 10 points). All other reviews are defined as the trusted reviews. Here, we exclude a single review written by a reviewer because it is classified as
Figure 6: A weakly supervised learning method for classifying movie reviews based on credibility.

trusted by the definition but cannot be judged.

As a comparison criterion for the proposed historical credibility, we choose the helpfulness vote. In general review systems, the helpfulness vote has been used as a meaningful measure for determining the value of the review, as explained in Section 2. Thus, Shan et al. and Yang et al. have shown that the helpfulness vote has a great influence on credibility evaluation [25, 26]. Further, Fang et al. have used the helpfulness vote as a criterion to evaluate the influence of various review factors on review credibility [24]. Similarly, with movie reviews, the helpfulness vote can be an effective criterion in determining review value.

Table 2 shows annotation methods for historical credibility as the proposed criterion and the helpfulness vote as the comparison criterion. For historical credibility, we define distrusted reviews by Definition 1 and all other reviews as trusted. For the helpfulness vote, we define each movie review as trusted when the number of helpful votes for the review exceeds the number of unhelpful votes.
Table 2: Data annotation based on credibility criteria

| Movie reviews | Historical credibility | Helpfulness vote [25, 26] |
|---------------|------------------------|---------------------------|
| Distrusted review | Review that is defined as distrusted by Definition 1 | Review with fewer helpful than unhelpful votes, or equal numbers of each kind of vote |
| Trusted review | Review that is defined as trusted by Definition 1 | Review with more helpful than unhelpful votes |

5 Performance Evaluation

5.1 Data collection

In this section, we describe the method we used to collect movie reviews and present the results of the data collection. We use BeautifulSoup and Selenium to implement a crawler to collect movie reviews from a movie website. The target movie website is Naver Movie\(^5\), the largest movie website in Korea\(^6\). From this website, we can collect all the historical movie reviews for a specific user; this is an essential condition for verifying the concept proposed in this paper. In addition, we collect the number of helpful or unhelpful votes for each movie review, as required for the comparison method. Only if these conditions are satisfied can we apply our concept into data sets crawled from any other website carrying movie reviews.

We consider various movie genres, comparing learning for movie reviews within the same genre and across different genres. For this purpose, reviews of a total of five movies from three different genres were collected. We collect movie reviews having both movie ratings and textual reviews. The movie

---

\(^5\)https://movie.naver.com/.

\(^6\)As of October 17, 2019, Naver was managing 12,539,719 ratings and 2,086,503 textual reviews.
ratings range from 1 through 10 points; the textual reviews are described in up to 140 characters. We also collect all the historical movie ratings for the movie reviewers included for the five movies chosen. The total number of historical movie ratings for these movie reviewers was 87,125. Table 3 shows the characteristics of the movie reviews collected. We collected 40,000 movie reviews, 8,000 for each movie. Here, 32,000 reviews are used for training and 8,000 are used for testing.

Table 3: Characteristics of collected movie reviews

| Genres  | Movies                | Release date | Training data | Testing data |
|---------|-----------------------|--------------|---------------|--------------|
| Drama   | Assassination         | 22 Jul. 2015 | 6,400         | 1,600        |
|         | I can speak           | 21 Sep. 2017 | 6,400         | 1,600        |
|         | The Spy Gone North    | 08 Aug. 2018 | 6,400         | 1,600        |
| Comedy  | Intimate Strangers    | 31 Oct. 2018 | 6,400         | 1,600        |
| Action  | Confidential Assignment| 18 Jan. 2017 | 6,400         | 1,600        |
| Total   |                       |              | 32,000        | 8,000        |

5.2 Experimental methods

In the experiments, we measure the elapsed learning time for movie reviews to check the efficiency of annotation in the proposed learning method. In addition, we measure the accuracy of the proposed learning method using historical credibility and that of the comparison method using the helpfulness vote as the classification criterion. Here, we measure the accuracy of combinations of the text representation models (TF-IDF and Word2Vec) and machine learning techniques (naïve Bayes and SVM). For TF-IDF and Word2Vec, we first need to extract keywords from the movie reviews. For this, we use a Korean keyword extractor\(^7\) and select the top 20 keywords based on frequency for each movie review. In TF-IDF, we apply a logarithm to TF to smooth its significance because the same proper nouns, such as actor/actress and movie names, tend to be repeated in movie reviews. In Word2Vec, we use a skip-gram

---

\(^7\)https://pypi.org/project/python-mecab-ko/
to consider the relevance of given keywords such as actor/actress and movie names in the movie reviews. We use a Gaussian kernel function for SVM, which has been widely used to classify textual data on high dimensions [38]. For the experiment, the entire data set is annotated according to the annotation methods presented in Section 4.2.

The experimental results consist of three parts:

1. Elapsed time for the proposed method

2. Accuracy comparison between the proposed method and the helpfulness vote

3. Variation in accuracy of the proposed method

In the first part, we show the efficiency of the proposed learning method. In the second part, we use the following experimental variables: 1) textual representation models, 2) machine learning techniques, and 3) variety of movie genres. In the third part, we use the following experimental variables: 1) number of movies, 2) machine learning techniques, and 3) textual representation models.

5.3 Experimental results

Elapsed time for the proposed method

Table 4 shows the average elapsed learning time for the five chosen movies using the proposed learning method using a combination of text representation models and machine learning techniques. Here, to measure the elapsed time, we use 8,000 movie reviews for each movie. We note that annotation of 8,000 movie reviews requires only 0.712 seconds on average. In particular, we note that when we use Word2Vec as the text representation model, which shows better performance than TF-IDF as will be shown in Section 5.3.3, the portion of the entire training process occupied by annotation is only about 0.145% ∼ 0.280%. By comparison, manual annotation to judge the credibility of 8,000 movie reviews requires much more time in absolute terms; the helpfulness vote approach also requires much time to collect enough votes after the movie reviews are written.
Table 4: Elapsed learning time for movie reviews (seconds)

| Combination     | TF-IDF + NB | W2V + NB | TF-IDF + SVM | W2V + SVM |
|-----------------|-------------|----------|--------------|-----------|
| Annotation      | 0.712       |          |              |           |
| Training        | 0.122       | 195.822  | 3.533        | 288.114   |
| Testing         | 0.040       | 57.615   | 0.027        | 200.070   |
| Total           | 0.874       | 254.149  | 4.273        | 488.896   |

5.3.1 Accuracy comparison between the proposed method and the helpfulness vote

Fig. 7 compares the accuracy of the methods by credibility criteria, namely the proposed historical credibility and the helpfulness vote. Here, we measure the accuracy for each combination of text representation models, namely TF-IDF or Word2Vec, and machine learning techniques, namely naïve Bayes or SVM. The results indicate that the accuracy of the proposed historical credibility is comparable to or even better than that of the helpfulness vote for all combinations of text representation models and machine learning techniques.

Fig. 7a shows the accuracy when we use TF-IDF as the text representation model and naïve Bayes as the machine learning technique. The result shows that mean accuracy for the helpfulness vote is 0.506, while that for historical credibility is 0.526. That is, the proposed historical credibility shows better accuracy than the helpfulness vote by a margin of 3.92%. Fig. 7b shows the accuracy when we use TF-IDF with SVM as the machine learning technique. The result shows that the proposed historical credibility has an accuracy improved over the helpfulness vote by 4.41%. Fig. 7c shows the accuracy when we use Word2Vec as the text representation model and naïve Bayes as the machine learning technique. The result shows that the proposed historical credibility has an accuracy improved over the helpfulness vote by 1.55%. Fig. 7d shows that the accuracy of the proposed historical credibility when we use Word2Vec and SVM is improved over the accuracy of the helpfulness vote by 2.49%.

Fig. 8 compares the accuracy of the methods by credibility criteria, namely historical credibility and helpfulness vote, and by variety of movie genres. Here, we use three movies having the same genre, drama, and then having different genres: drama, comedy, and action. The result shows that in
Figure 7: Accuracy comparison by combination of text representation models and machine learning techniques

Improvement in accuracy of the proposed historical credibility over that of the helpfulness vote in both cases. Specifically, the proposed historical credibility shows accuracy improvements of 1.57% in the case of the same genre and 4.54% in the case of the three different genres.
Variation in accuracy of the proposed method

Fig. 9 shows the accuracy of the proposed historical credibility by text representation model, namely TF-IDF and Word2Vec. Here, we select three movies from different genres: Assassination, Intimate Strangers, and Confidential Assignment. Fig. 9a shows the accuracy of TF-IDF and that of Word2Vec when we keep the machine learning technique fixed as naïve Bayes; Fig. 9b shows the respective accuracy with the machine learning technique fixed as SVM. Both results show better accuracy when using Word2Vec than when using TF-IDF. Specifically, for the naïve Bayes learning technique, mean accuracy using TF-IDF is 0.526 compared with 0.552 using Word2Vec; in the SVM technique, the mean accuracy using TF-IDF is 0.529 compared with 0.551 using Word2Vec.

Fig. 10 shows the accuracy by machine learning technique, namely naïve Bayes and SVM. Here, we select the same three movies from different genres as before. Fig. 10a shows the accuracy of naïve Bayes and that of SVM when we keep the text representation model fixed as TF-IDF. Fig. 10b shows the respective accuracy with the text representation model fixed as Word2Vec. The results show slightly better overall accuracy when using SVM than when using naïve Bayes. Specifically, in TF-IDF, mean accuracy using naïve Bayes is 0.526 compared with 0.529 using SVM. In Word2Vec, mean accuracy
Figure 9: Accuracy variation by text representation model

(a) Naïve Bayes
(b) SVM

Figure 10: Accuracy variation by machine learning technique

(a) TF-IDF
(b) Word2Vec

6 Discussions

Language-independent method: In this paper, we use movie reviews collected from the largest Korean movie review website, Naver Movie, because it satisfies two requirements to show the effectiveness
of the proposed method: 1) it maintains all the historical movie reviews written by each user, as required for the proposed historical credibility method; 2) it maintains helpful and unhelpful votes for each review, as required for the comparison method based on the helpfulness vote. We can apply and test the proposed method to movie reviews collected from other websites only if they satisfy the two requirements above.

We note that the proposed method is language-independent. The accuracy of the learning model might vary according to the language used in movie reviews. In this paper, we have focused on the relative accuracy difference between methods, not on the absolute value of the accuracy for the target language. This relative difference should be maintained in different languages.

**Variation in accuracy according to text representation model and machine learning technique:** In this paper, we aim to show the effectiveness of the proposed learning method, rather than the superiority of a specific text representation model or a specific machine learning technique. For this reason, we choose two representative text representation models and two representative machine learning techniques. However, we could apply any other text representation model and/or machine learning technique to the proposed method.

### 7 Conclusions

In this paper, we have proposed a weakly supervised learning method for judging the credibility of movie reviews. The proposed learning method has two advantages. First, it is efficient because it allows fast annotation according to predefined rules. Second, it is effective because it uses a simple but effective criterion to judge movie review credibility based on historical ratings.

For the experiments, we collected 40,000 actual movie reviews. We have shown that the proposed learning method is efficient because it can annotate 8,000 movie reviews in only 0.712 seconds, which occupies only about 0.145% of the entire training process when we use Word2Vec as the text representation model. We have also measured the accuracy of the proposed learning method. By comparing accuracy between the proposed learning method and the comparison method based on the helpfulness vote, which has been used as an effective criterion in many studies, we have shown that the accuracy
of the proposed method is better than that of the comparison method by 1.57% ∼ 4.54% depending on the text representation model, machine learning technique, and movie genres used.

References

[1] Chintagunta PK, Gopinath S, Venkataraman S. The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. Marketing Science. 2010;29(5):944–957.

[2] Dong XL, Rekatsinas T. Data integration and machine learning: A natural synergy. In: Proceedings of the 2018 International Conference on Management of Data. ACM; 2018. p. 1645–1650.

[3] Raykar VC, Yu S, Zhao LH, Valadez GH, Florin C, Bogoni L, et al. Learning from crowds. Journal of Machine Learning Research. 2010;11(Apr):1297–1322.

[4] Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics; 2009. p. 1003–1011.

[5] Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: Rapid training data creation with weak supervision. Proceedings of the VLDB Endowment. 2017;11(3):269–282.

[6] Ratner AJ, De Sa CM, Wu S, Selsam D, Ré C. Data programming: Creating large training sets, quickly. In: Advances in neural information processing systems; 2016. p. 3567–3575.

[7] Lee G, Jeong J, Seo S, Kim C, Kang P. Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network. Knowledge-Based Systems. 2018;152:70–82.

[8] Lin D, Li L, Cao D, Lv Y, Ke X. Multi-modality weakly labeled sentiment learning based on Explicit Emotion Signal for Chinese microblog. Neurocomputing. 2018;272:258–269.
[9] Serrano-Guerrero J, Olivas JA, Romero FP, Herrera-Viedma E. Sentiment analysis: A review and comparative analysis of web services. Information Sciences. 2015;311:18–38.

[10] Topal K, Ozsoyoglu G. Movie review analysis: Emotion analysis of IMDb movie reviews. In: Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE Press; 2016. p. 1170–1176.

[11] Manek AS, Shenoy PD, Mohan MC, Venugopal K. Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. World wide web. 2017;20(2):135–154.

[12] Chakraborty K, Bhattacharyya S, Bag R, Hassanien AE. Comparative sentiment analysis on a set of movie reviews using deep learning approach. In: International Conference on Advanced Machine Learning Technologies and Applications. Springer; 2018. p. 311–318.

[13] Alsaqer AF, Sasi S. Movie review summarization and sentiment analysis using rapidminer. In: 2017 International Conference on Networks & Advances in Computational Technologies (NetACT). IEEE; 2017. p. 329–335.

[14] He Y, Zhou D. Self-training from labeled features for sentiment analysis. Information Processing & Management. 2011;47(4):606–616.

[15] Elmurngi E, Gherbi A. Fake Reviews Detection on Movie Reviews through Sentiment Analysis Using Supervised Learning Techniques. International Journal on Advances in Systems and Measurements. 2018;11(1 & 2):196–207.

[16] Lee HA, Law R, Murphy J. Helpful reviewers in TripAdvisor, an online travel community. Journal of Travel & Tourism Marketing. 2011;28(7):675–688.

[17] Mudambi S, Schuff D. What Makes a Helpful Review? A Study of Customer Reviews on Amazon.com (SSRN Scholarly Paper No. ID 2175066). Social Science Research Network, Rochester, NY. 2010;.
[18] Ham J, Lee K, Kim T, Koo C. Subjective perception patterns of online reviews: A comparison of utilitarian and hedonic values. Information Processing & Management. 2019;56(4):1439–1456.

[19] Liu Z, Park S. What makes a useful online review? Implication for travel product websites. Tourism Management. 2015;47:140–151.

[20] Ghose A, Ipeirotis PG. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. IEEE Transactions on Knowledge and Data Engineering. 2010;23(10):1498–1512.

[21] Hochmeister M, Gretzel U, Werthner H. Destination expertise in online travel communities. In: Information and Communication Technologies in Tourism 2013. Springer; 2013. p. 218–229.

[22] Barbado R, Araque O, Iglesias CA. A framework for fake review detection in online consumer electronics retailers. Information Processing & Management. 2019;56(4):1234–1244.

[23] Reyes-Menendez A, Saura JR, Martinez-Navalon JG. The impact of e-WOM on hotels management reputation: exploring tripadvisor review credibility with the ELM model. IEEE Access. 2019;7:68868–68877.

[24] Fang B, Ye Q, Kucukusta D, Law R. Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. Tourism Management. 2016;52:498–506.

[25] Shan Y. How credible are online product reviews? The effects of self-generated and system-generated cues on source credibility evaluation. Computers in Human Behavior. 2016;55:633–641.

[26] Yang SB, Shin SH, Joun Y, Koo C. Exploring the comparative importance of online hotel reviews' heuristic attributes in review helpfulness: a conjoint analysis approach. Journal of Travel & Tourism Marketing. 2017;34(7):963–985.

[27] Ren G, Hong T. Examining the relationship between specific negative emotions and the perceived helpfulness of online reviews. Information Processing & Management. 2019;56(4):1425–1438.

[28] Jones KS. A statistical interpretation of term specificity and its application in retrieval. Journal of documentation. 2004;.
[29] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems; 2013. p. 3111–3119.

[30] Wu HC, Luk RWP, Wong KF, Kwok KL. Interpreting tf-idf term weights as making relevance decisions. ACM Transactions on Information Systems (TOIS). 2008;26(3):13.

[31] Chowdhury GG. Introduction to modern information retrieval. Facet publishing; 2010.

[32] Ling W, Dyer C, Black AW, Trancoso I. Two/too simple adaptations of word2vec for syntax problems. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2015. p. 1299–1304.

[33] Ikonomakis M, Kotsiantis S, Tampakas V. Text classification using machine learning techniques. WSEAS transactions on computers. 2005;4(8):966–974.

[34] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics; 2002. p. 79–86.

[35] Ch R, MAP M. Bayesian learning. Machine Learning. 1997;.

[36] Vapnik V. The nature of statistical learning theory. Springer science & business media; 2013.

[37] Wilk-Kolodziejczyk D, Regulski K, Gumienny G. Comparative analysis of the properties of the nodular cast iron with carbides and the austempered ductile iron with use of the machine learning and the support vector machine. The International Journal of Advanced Manufacturing Technology. 2016;87(1-4):1077–1093.

[38] Keerthi SS, Lin CJ. Asymptotic behaviors of support vector machines with Gaussian kernel. Neural computation. 2003;15(7):1667–1689.