12-7-2022

Are Online Consumer Reviews Credible? A Predictive Model based on Deep Learning

Ehsan Abedin
The University of Melbourne, eabedin@student.unimelb.edu.au

Antonette Mendoza
The University of Melbourne, mendozaa@unimelb.edu.au

Shanika Karunasekera
The University of Melbourne, karus@unimelb.edu.au

Are Online Consumer Reviews Credible? A Predictive Model based on Deep Learning

Full research paper

Ehsan Abedin
School of Computing and Information Systems
University of Melbourne
Victoria, Australia
Email: eabedin@student.unimelb.edu.au

Antonette Mendoza
School of Computing and Information Systems
University of Melbourne
Victoria, Australia
Email: mendozaa@unimelb.edu.au

Shanika Karunasekera
School of Computing and Information Systems
University of Melbourne
Victoria, Australia
Email: karus@unimelb.edu.au

Abstract

As the importance of online consumer reviews has grown, the concerns about their credibility being damaged by the presence of fake reviews have also grown. Extant literature reveals the importance of online reviews for consumers. Yet, there is a lack of research in the literature that considers consumer perception while developing a predictive model for the credibility of online reviews. This research aims to fill this gap by combining two different streams in the literature namely human-driven and data-driven approaches. To do so, we use two datasets with different labelling approaches to develop a predictive model, the first one is labelled based on the Yelp filtering algorithm and the second one is labelled based on the crowd's perception towards credibility. Results from our predictive model reveal that it can predict credibility with a performance of 82% AUC, using reviews' attributes namely, length, subjectivity, readability, extremity, external and internal consistency.

Keywords Online Consumer Reviews, Review Credibility, Deception Detection, Fake Reviews.

1 Introduction

Online reviews play an important role in consumers' purchasing decisions about products or services. As the importance of online reviews has grown, the concerns about their credibility being damaged by the presence of fake reviews have also grown; businesses post promotional reviews for themselves or deceptive negative reviews for their competitors (Ansari and Gupta 2021; Luca and Zervas 2016). A recent report has shown that more than 200K people are involved in posting fake reviews only on Amazon website[1]. In addition, Yelp itself acknowledged that up to 25% of online reviews on its website are at least suspicious[2,3]. To tackle this problem, prior research (e.g., Ansari and Gupta 2021; Barbado et al. 2019; Jindal and Liu 2008; Mukherjee et al. 2013a; Mukherjee et al. 2013b; Ott et al. 2013; Ott et al. 2011; Rayana and Akoglu 2015; Salminen et al. 2022; Zhang et al. 2016) has attempted to develop models to filter out fake reviews from online review websites, like Amazon, Yelp, Expedia, TripAdvisor, and Flipkart. These studies and tools are useful; however, they have their own limitations.

Firstly, most previous studies have used information about the source of a review (i.e., reviewer) to assess the credibility of their written reviews. Although information about the source of a review is found to be helpful in detecting fake reviews (Wu et al. 2020), many online review platforms (e.g., Yelp), do not allow anyone to collect or process information about their users due to their terms and conditions, privacy and ethics issues. More importantly, many online review platforms (e.g., insureye.com, ratemds.com) do not capture and accordingly present information about their users. Thus, information about the source of a review is not always accessible[4]. Therefore, it is important to understand to what extent we can predict the credibility of online reviews without using information about its source and, instead, considering other important information. Particularly, this is vital for small e-commerce websites or online review aggregators that do not have any models or tools to assess the credibility of their online reviews and do not collect information about their users. Also, there is information such as the consistency of a single review with other reviews (i.e., external consistency) and the consistency of different elements within a review (i.e., internal consistency) - that have not received enough attention from previous studies, which we investigate as a predictor for the credibility of online reviews.

Secondly, there is a lack of research in the literature that considers consumer perception while developing predictive models for the credibility of online reviews and differentiates the characteristics of fake and credible reviews using attributes that consumers actually employ while assessing the credibility of online reviews. We believe, consumers are the main user of online reviews, and this information is written to help consumers make a better purchasing decision about a product or service; accordingly, it would not be highly useful to develop a predictive model for consumers without considering their perspective. Thus, based on literature, we, first, define credibility as "believability" or "the characteristic that makes people trust and believe something or someone"(Cheung et al. 2012); then, we answer the following research question to address the above-mentioned limitations in the literature: *To what extent can we predict the credibility of online reviews without using information about its source?*

The contributions in the paper help consumers make better purchasing decisions based on credible information by realizing the differences between fake and credible reviews. It also assists e-commerce platforms to measure the credibility of their online reviews and provide reliable information for their users. Particularly, this is vital for small e-commerce websites or online review aggregators that do not have any models or tools to filter out fake/suspicious reviews and do not collect information about their users (e.g., insureye.com, ratemds.com). The remainder of this paper is organized as follows. We present our literature review around the credibility evaluation of online consumer reviews in the next section. This is followed by a discussion about the research methodology of this study to address the research question. Next, we present the results of this research including our predictive model. Finally, we explain the discussion and contributions of this research.

2 Research Background

Evaluating the credibility of online reviews and identifying fake reviews is a particular application of the general issue of deception detection (Barbado et al. 2019; Fitzpatrick et al. 2015). According to our

[1] https://au.pcmag.com/shopping/87059/database-reveals-over-200k-people-involved-in-posting-fake-reviews-on-amazon
[2] https://fortune.com/2013/09/26/yelps-fake-review-problem/
[3] https://www.yelp-press.com/company/fast-facts/default.aspx
[4] Although this information might be available for a website administrator, it is not always accessible for a third party like researchers. For instance, Yelp platform, which has been used in this work, does not permit us to "record, process, or mine information about its users" and has mentioned this in its website.

literature review, Jindal and Liu (2007) were the first researchers who tried to assess the credibility of online reviews. This work analyzed online reviews on Amazon website and considered duplicates or nearly duplicates reviews as fake to develop a model that identifies suspicious reviews. The authors of this work argued that fake reviews are much more difficult to identify than other deception detection problems on the internet. After this work, several scholars attempted to tackle the problem of fake reviews from different perspectives (Wu et al. 2020). Based on our literature analysis, we can classify studies regarding the credibility of online reviews into two main categories: (1) studies with human-driven approaches and (2) studies with data-driven approaches.

The first stream (i.e., human-driven) is based on developing a theoretical model and carefully exploring the attributes that impact the credibility of online reviews by conducting user studies and incorporating theories from different domains such as psychology, communication, and information systems to support their findings (Abedin et al. 2021). Examples of this stream can be quantitative research (e.g., experiment or questionnaire) or qualitative research (e.g., interview). For instance, Cheung et al. (2012) investigated the impacts of review sidedness, source credibility, review consistency and argument quality on the credibility assessment of online reviews, using the elaboration likelihood model (ELM) as their theoretical lens. In another study, Filieri (2016) used a grounded theory method to explore how travellers judge online reviews' trustworthiness. This study conducted 38 interviews with travellers on TripAdvisor and proposed a theoretical model to explain how these travellers assess review trustworthiness. Abedin et al. (2019b) developed a conceptual model and identified several attributes (e.g., review length, source credibility and consistency among reviews) that impact the credibility of online reviews by conducting 21 interviews with online shoppers and using the Heuristic Systematic Model (HSM) as a theoretical lens.

The second stream (i.e., data-driven) is built on developing machine learning models including supervised (Barbado et al. 2019; Kumar et al. 2018; Liu et al. 2019), semi-supervised (Rout et al. 2017a; Rout et al. 2017b; Zhang et al. 2017) and unsupervised techniques (Dong et al. 2018) to automate fake reviews detection. For example, Kumar et al. (2018) proposed a hierarchical supervised-learning model by considering several users' attributes to improve the likelihood of detecting fake reviews. Using logistic regression, they could reach an Area Under the Curve (AUC) score of 0.817. Plotkina et al. (2020) explored the characteristics of fake reviews. The authors tried to combine multiple-micro linguistic cues to develop an algorithm that can identify fake reviews. Their findings show that fake reviews have fewer paragraphs, contain more adjectives than verbs, are structured and easier to understand. Barbado et al. (2019) proposed a feature framework to detect fake online reviews that has been evaluated in the consumer electronics domain. Using the Ada Boost classifier, they achieved an 82% F-Score on the detection of fake reviews. Lu et al. (2013) took a different approach which tried to detect fake reviewers and fake reviews at the same time. To do so, the authors proposed a Review Factor Graph model which incorporates several review and reviewer attributes.

As can be seen, the first stream mostly focused on identifying the important attributes from consumers perspective that impact the credibility of online reviews; however, these studies have not developed a predictive model to explore to what extent we can predict the credibility of online reviews using consumers' perspective. On the other hand, studies in the second stream mainly attempted to develop machine learning models to filter out fake reviews and improve the performance of their models. Nonetheless, these studies mainly do not consider consumers perspective in the development of their models. In this study, we take advantage of both streams and address their limitations by combining both approaches, applying attributes that consumers use, studied in the human-driven approach and employing predictive models, used in the data-driven approach. In addition, to better incorporate human-driven approach into the data-driven approach, on top of the Yelp dataset and using its filtering algorithm as the target label, we collect another set of data with a different labelling approach. To do so, we consider the crowd's perception towards the credibility of online reviews as a label and use it in the classifier for the credibility prediction of online reviews. In the following section, we explain the research method of this study including further explanation of two datasets used in this research.

3 Research Method

This section provides details about the steps we carry out in this particular study to develop the predictive model for the credibility of online reviews.

3.1 Attributes Used for the Development of the Predictive Model

As discussed in the "research background" section, some studies have investigated the credibility of online reviews from consumers' perspective (e.g., Abedin et al. 2019b; Filieri 2016; Luo et al. 2015).

Although these studies have identified attributes that consumers use while assessing online reviews, they still have not investigated the extent to which we can predict the credibility of online reviews by using these attributes. Thus, in this study, we use attributes from these studies that can be used to develop a predictive model for the credibility of online reviews. These attributes are: review extremity, review length, external consistency, review subjectivity, internal consistency, review readability (fluency) and source of a review. However, as discussed in the introduction section, due to privacy, ethics, terms and conditions issues, we do not use information about the source of a review in our study.

Figure below visualizes these attributes and steps we carry out to extract these attributes to develop the predictive model.

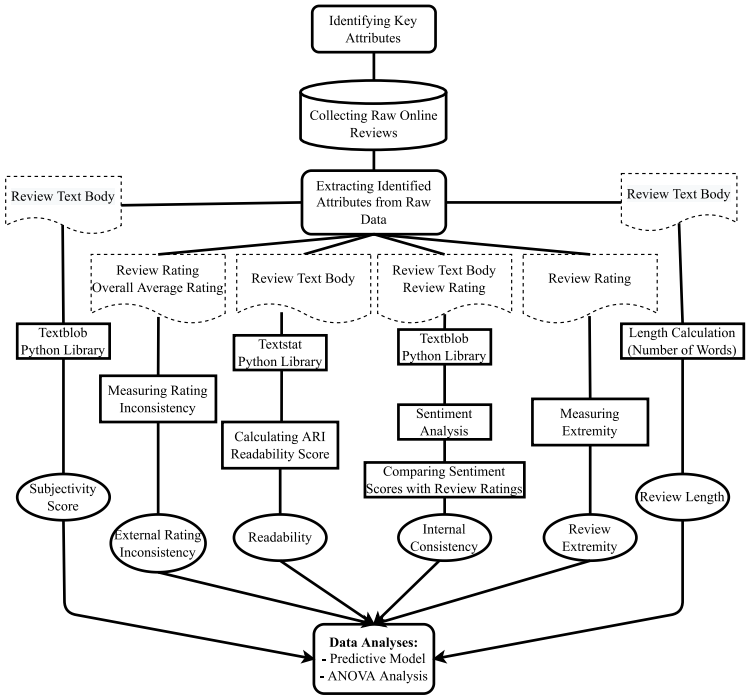

Figure 1: Attributes Used for the Predictive Model

The length of each review is calculated, by counting the number of words in its textual content. Review extremity is measured as the star rating of the review. A very low rating (i.e., one star) or a very high rating (i.e., five stars) indicate extreme reviews and a review with a star rating of 2,3 or 4 reflects a non-extreme review. Rating inconsistency (external consistency) is calculated as the absolute difference between a review's star rating and the overall average rating for a business (Aghakhani et al. 2020).

Next, to extract review subjectivity, we use Textblob[5] to perform subjectivity analysis on each reviews' textual content. Textblob is a Python Natural Language Processing (NLP) library for processing textual data. Subjectivity scores of reviews vary between 0.0 and 1.0, where 0.0 presents a very objective review and 1.0 reflects a very subjective one (Loria 2018). Internal consistency refers to the consistency within a single review, which is the consistency between the content of a review and its valence (e.g., stars rating). Thus, if the content and stars rating are not aligned, the review is internally inconsistent; otherwise, the review is consistent. For instance, if the text body of a review shows a positive opinion (positive sentiment) but the review has a low star rating (e.g., 1 or 2 out of 5), it shows that this review is internally inconsistent. On the other hand, if the text body is written positively and the star rating of the review is also high (e.g., 4 or 5 out of 5), it shows that the review is internally consistent (Abedin et al. 2019a; Abedin et al. 2020) . In order to extract this attribute, we first conduct the sentiment analysis on each review text body, using the Textblob library. Then we compare the sentiment score for each review with its stars rating. If the sentiment score and stars rating are opposites and do not match, the review is internally inconsistent; otherwise, it is internally consistent. Finally, to extract review readability (fluency), we use Textstat[6] Python NLP library to calculate the Automated Readability Index (ARI) for each review's textual content. We have chosen ARI index in this study as ARI is one of the main

[5] https://textblob.readthedocs.io/en/dev

[6] https://pypi.org/project/textstat

readability tests that were used to assess the readability of a text and it is subjected to a lower error rate as compared to other readability measures (Hu et al. 2011).

ARI is calculated using its standard formula (Senter and Smith 1967):

$$ARI\ Score = 4.71\left(\frac{number\ of\ characters}{number\ of\ words}\right) + 0.5\left(\frac{number\ of\ words}{number\ of\ sentences}\right) - 21.43$$

Table below summarizes descriptions of each attribute along with their possible values.

Attribute Name	Description	Possible Values
Review Credibility	Is the review recommended or filtered?	0 = Filtered, 1 = Recommended
Review Length	Number of words in a review	0 – No upper limit
Review Extremity	Is the review an extreme one?	0 = No, 1 = Yes
External Rating Inconsistency	|a review's star rating - overall average rating|	0 - 4
Review Subjectivity	The subjectivity score of a review from Textblob library	0 - 1
Internal Consistency	Does the review have consistency between its star ratings and its sentiment score?	0 = No, 1 = Yes
Review Readability	The readability score of a review from Textstat library	No upper or lower limit

Table 1. Attributes Description

3.2 Data Collection

The aim of this study is to develop a predictive model to measure the credibility of online reviews by using attributes that consumer employs (i.e., review extremity, review length, external consistency, review subjectivity, internal consistency and review readability/fluency). To do so, we use two datasets from Yelp platform to get a better understanding around the importance of these attributes in two conditions and realize to what extent these attributes can predict the credibility of online reviews. The first dataset is labelled based on Yelp filtering algorithm and the second one is labelled based on the crowd's perception towards the credibility of online reviews. In the followings, we further explain these two datasets including the reason behind why we have selected Yelp platform as the case in this study and why we use two datasets with different labelling approaches.

3.2.1 Dataset One (Considering Yelp Filtering Algorithm)

To create our first dataset, we have collected 12000 reviews of services from the Yelp review dataset (Yelp dataset challenge) used by (Barbado et al. 2019; Mukherjee et al. 2013b; Rayana and Akoglu 2015), through a random selection to conduct our analyses and develop the predictive model for the credibility evaluation of online reviews. We have considered Yelp platform due to the following reasons. First, Yelp dataset[7] has been widely used by previous researchers (Barbado et al. 2019; Luca and Zervas 2016; Mukherjee et al. 2013b; Rayana and Akoglu 2015; Zhang et al. 2016). A key reason for using Yelp is that it uses a filtering algorithm that identifies suspicious/fake reviews and puts them into a filtered list. According to Yelp's CEO, its filtering algorithm has evolved over the years to filter out suspicious reviews. In addition, its filtering algorithm has been claimed to be highly accurate by previous researchers (Barbado et al. 2019; Mukherjee et al. 2013b; Rayana and Akoglu 2015).

Second, although Yelp does not reveal how its algorithm works, Yelp is confident enough to make both suspicious/filtered and recommended reviews public, and as such, it is possible to view the reviews that are not currently recommended through a link at the bottom of a business page. With this in hand, we are able to analyze the patterns of fake and credible reviews on Yelp (Barbado et al. 2019; Mukherjee et al. 2013b; Rayana and Akoglu 2015; Weise 2011). Thus, in this study, we use online consumer reviews from Yelp to create our first dataset.

3.2.2 Dataset Two (Considering Crowds Perception)

Yelp dataset provides us with online reviews as data. In addition, as mentioned before, it enables us to see whether an online review has been filtered or not. However, this labelling process of online reviews (i.e., recommended or not) is like a black box as Yelp model/system does not show how this assessment of online reviews has been occurred. For instance, it is not clear whether Yelp model is purely computer-

[7] Sometimes known as Yelp Dataset Challenge.

based, human-based or something between. Thus, to further incorporate users' perspectives towards the development of the predictive model for the credibility of online reviews and further validate the importance of the attributes namely, review extremity, review length, external consistency, review subjectivity, internal consistency and review readability (fluency) in this process, we conduct an online user study to employ the crowd's opinion as the label for the credibility of online reviews. Thus, for the second dataset, we randomly select 200 reviews as data from a restaurant on Yelp.
To label this dataset based on the crowd's evaluation, in the first step, we shared the experiment's link with the potential participants on a social media platform (i.e., Linkedin) and a group of students at a university in Australia via their emails. During the data collection process, all the respondents were notified that they would receive a 40\$ e-gift voucher for their participation. If they agreed, we provided them with the experiment's link. Finally, we asked the participants to recommend other potential participants as the snowballing sampling technique. At the end of the data collection process, after removing incomplete and partial responses, in total, we collected 9800 judgments for 200 reviews in the dataset two; meaning that each review is rated by 49 participants on average. We designed the experiment in a way that captures the dependent variable – credibility, for each of the reviews in a Likert scale instead of a binary scale. This is because Likert scale provides us with a more granular result as it gives participants more degrees of opinions to choose, and make their decisions, rather than confining them with only two options. In addition, after collecting all data, the result is a scale mean, which shows the collective opinion of crowd towards each review's credibility, which can easily be converted to either fake or credible, based on introducing a threshold (Spooren et al. 2007; Wu and Leung 2017).

Among different Likert scales, we have chosen the Likert scale from 0 (most fake) to 10 (most credible) as a natural and easily comprehensible range, recommended by Leung (2011), Hodge and Gillespie (2007) and Wu and Leung (2017). As stated by Leung (2011), Likert scale from 1 to 10 "increases sensitivity and is closer to interval level of scaling and normality". Thus, in this study, we first use the Likert scale from 0-10 to get the opinions of crowds towards the credibility of a given review. Next, we aggregate these credibility scores, rated by participants to calculate the mean value for each review. Finally, based on the mean value and introducing a cut-off point, we label each review as credible or fake. In order to find an appropriate cut-off point, we did the ROC analysis based on the calculated average score and the Yelp labels. To do so, we examined different thresholds of: 4.5, 5, 5.5, 6 and 6.5. Table 2 present the result of our ROC analysis.

Test Result Variable(s)	Area	Std. Error	Asymptotic Sig.
Cut off point: 4.5	.625	.023	.000
Cut off point: 5	.865	.023	.000
Cut off point: 5.5	.840	.026	.000
Cut off point: 6	.801	.028	.000
Cut off point: 6.5	.673	.031	.000

Table 2. Area Under the ROC Curve for Different Cut-off Points

As can be seen from Table 2, the cut-off point of 5 has the best performance in terms of the area under the ROC. Thus, we labelled those reviews that are below the threshold of 5 as fake (0) and those above the threshold as credible (1).

3.3 Selecting the Predictive Model

To select an appropriate machine learning model to develop our predictive model, we have tested a variety of algorithms including Decision Tree (DT), Random Forest (RF), K Nearest Neighbours (KNN), Support Vector Machine (SVM), Logistic Regression (LR) and Artificial Neural Network (ANN) – deep learning – to realize which model presents the best performance regarding the credibility prediction of online reviews. As will be discussed further in the findings section (Table 4 and Table 5), among all those models, deep learning showed the best performance in this research. Thus, we use Artificial Neural Network (deep learning) to create our predictive model and evaluate the credibility of a given review. Thus, in what follows, we discuss deep learning in further detail.

3.3.1 Deep Learning

Deep learning (also known as deep machine learning or deep structured learning) is a branch of machine learning, based on a set of dynamic systems. It consists of a set of artificial neurons which are connected by a synapse that try to make a linear or non-linear mapping from a group of inputs to target output/s (Basheer and Hajmeer 2000; Eslami et al. 2018; LeCun et al. 2015).

According to LeCun et al. (2015), "Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction". Thus, this computational power makes it a very practical tool to address a variety of problems, such as object detection, speech recognition and many other areas such as genomics and drug discovery [8] (LeCun et al. 2015). Among all deep learning architectures, multi-layer perceptron model has been chosen and used in this research as it is the most prominent one (Bischof et al. 1992; Eslami et al. 2018).

The first step in creating a deep learning model is to choose the right architecture by finding the appropriate number of inputs, hidden layers, outputs and the number of neurons in each layer (Basheer and Hajmeer 2000; Eslami et al. 2018; LeCun et al. 2015). We call the first layer as the input layer, the last layer as the output layer and finally layer/s between the input and output layers as the hidden layer/s. In this study, we used review extremity, review subjectivity, review readability, internal consistency, review length, and rating inconsistency (external consistency) to predict the credibility of online reviews. Thus, our input layer is made up of six neurons signifying these attributes, also the output layer is made up of one neuron signifying the credibility of online reviews.

Regarding the hidden layer/s, generally, there is no rule to calculate the number of layers or nodes per layer; however, there are some rules of thumb to create a model architecture. The first one is to keep the architecture as simple as possible (Eslami et al. 2018; Krimpenis et al. 2006; LeCun et al. 2015). Secondly, there is a theoretical finding by Goodfellow et al. (2017) and Heaton (2008) that shows two hidden layers are adequate for creating a classification model. Thirdly, as suggested by Heaton (2008), the number of hidden neurons in each hidden layer should be less than twice the size of the input layer. Thus, we use these guidelines and rules of thumb to develop our prediction model in the next section.

4 Findings

This section presents the predictive model, its result and its performance for the first and second dataset.

4.1 Results of The Predictive Model in the First Dataset

The first dataset contains 12000 online reviews from Yelp, which has been labelled based on its filtering algorithm. To develop the predictive model, we first extract the discussed attributes (i.e., review extremity, review length, external consistency, review subjectivity, internal consistency and review readability/fluency) from this dataset. Summary statistics of these attributes are shown in Table below.

Attribute Name	Mean	Median	Min	Max	Standard Deviation
Review Extremity	0.516	1.000	0.000	1.000	0.500
Rating Inconsistency (External)	0.926	1.000	0.000	3.500	0.786
Review Length	123.076	92.000	1.000	966.000	109.439
Review Subjectivity	0.575	0.566	0.000	1.000	0.115
Review Readability	20.031	10.400	-7.500	412.500	24.287
Internal Consistency	0.889	1.000	0.000	1.000	0.314

Table 3. Attributes Descriptive Statistics for the First Dataset (n = 12000)

Next, we use these attributes as the input to the first layer of our deep learning model. Afterwards, according to the guidelines and rules of thumb discussed in the research method section regarding how to create a multi-layer perceptron model, in the beginning, we started with two hidden layers and then we tried to change the number of nodes in each hidden layer to see if the model's performance improves or not. Ultimately, the deep learning model with 4 layers (input layer with 6 nodes, first hidden with 10 nodes, second hidden layer with 8 nodes, and output layer with one node) was found to have the best performance in our dataset. Next, the model was trained using 60% of our data as a training set, leaving 20% as the validation set and finally 20% as the test set. The training was discontinued when performance stopped to improve on the validation set. Finally, we assessed the performance of the model on our test set using the AUC as suggested by Norton and Uryasev (2019). Table 4 summarizes the performance of our model. As mentioned in the research method section, in addition to the discussed deep learning model, we employed a variety of other machine learning models namely: Decision Tree, Random Forest, K Nearest Neighbours, Support Vector Machine, and Logistic Regression to investigate which model has the best performance in predicting the credibility of a given review.

[8] Please refer to LeCun et al. (2015) for more information and a nice overview of deep learning.

	Performance (AUC)
Train Set	0.82
Validation Set	0.83
Test Set	0.82

Table 4. Our Deep Learning Model Performance

As shown in Table4 and Table 5, our deep learning shows a better performance in comparison with other machine learning models in our research. Thus, in this study, we developed our predictive model based on deep learning for the credibility evaluation of a given review.

Model	**Performance (AUC)**
Logistic Regression (LR)	0.79
K Nearest Neighbours (KNN)	0.78
Decision Tree (DT)	0.67
Random Forest (RF)	0.78
Support Vector Machine (SVM)	0.73

Table 5. Performance of Other Machine Learning Models

In order to increase the explainability of our findings, we attempted to present the predicted output of our deep learning model in a way that humans can understand, by visualizing the outputs using a pivot table-heatmap as illustrated in Figure 2. To do so, first, we checked the prediction power of our input

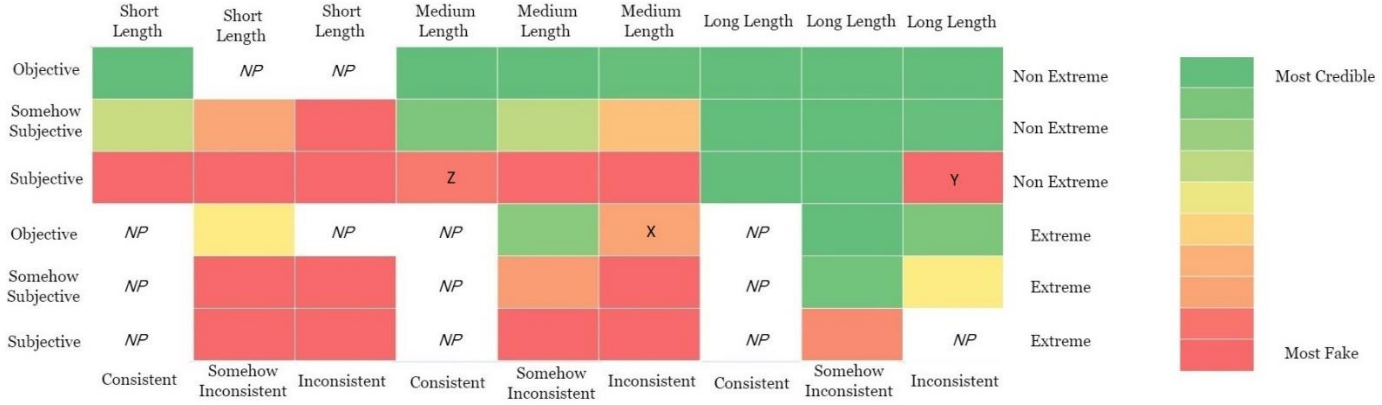

Figure 2: Our Deep Learning Model Prediction Results

attributes in our model. Among all the attributes, review length, review subjectivity and rating inconsistency (external) were found to be the strongest predictors of the credibility of online reviews followed by review extremity, review readability and internal consistency. Next, to better visualize our model output, we selected the top 4 predictors: review length, review subjectivity, rating inconsistency (external) and review extremity. Then, we did some classifications for these attributes as follows: regarding the review's length, we classified reviews that placed one standard deviation above or below the mean as long and short, respectively, and the rest were classified as medium length. Similarly, we did the same classifications for review subjectivity (i.e., subjective, somehow subjective and objective) and rating inconsistency (inconsistent, somehow inconsistent and consistent). Thus, it resulted in having three conditions for review length (short, medium and long), three conditions for review subjectivity (objective, somehow subjective and subjective), three conditions for reviews' rating inconsistency (consistent, somehow inconsistent and inconsistent) and finally two conditions for review extremity (extreme and non-extreme), leaving us with 54 different conditions (3* 3 * 3 * 2 = 54). Figure 2 shows the output of the predictive model for all these 54 conditions. As illustrated in Figure 2, the most credible reviews are those that are mainly associated with low subjectivity, long length, low rating inconsistency, and generally are among non-extreme reviews. On the other hand, the least credible reviews are those that are mostly associated with high subjectivity, low readability score, short length, high rating inconsistency, and usually are among extreme reviews.

4.2 Results of The Predictive Model in the Second Dataset

To further validate the predictive model and the attributes used as the input in the predictive model, we use our second dataset, which contains 200 reviews and has been labelled based on the crowd's

perception towards the credibility of reviews. In this regard, similar to the first dataset, we extract the related attributes from the second dataset. Summary statistics of these attributes are shown in Table 6.

Attribute Name	Mean	Median	Min	Max	Standard Deviation
Review Extremity	0.562	1.000	0.000	1.000	0.497
Rating Inconsistency (External)	1.268	1.000	1.000	3.000	0.598
Review Length	108.064	78.000	2.000	469.000	95.367
Review Subjectivity	0.588	0.594	0.000	1.000	0.146
Review Readability	17.380	9.100	-1.700	120.800	19.794
Internal Consistency	0.880	1.000	0.000	1.000	0.325

Table 6. Attributes Descriptive Statistics for the Second Dataset (n = 200)

According to the guideline presented in the research method and similar to the first dataset, we calculated the performance of the predictive model against the second dataset. Overall, our deep learning model presents the AUC between 80-82 for the second dataset, which is very similar to the first dataset (although slightly lower). These results validate the predictive model and confirm the importance of the attributes (i.e., review extremity, review length, external consistency, review subjectivity, internal consistency and review readability) used in the development of the model for the credibility evaluation of online reviews.

5 Discussion

In this study, we developed a predictive model based on deep learning that can evaluate the credibility of online reviews by incorporating consumers perspective and without using information about a source of a review to mitigate the problems arising from utilizing users' information. Instead, we employed some other attributes including internal and external consistency that have not received enough attention from previous studies in the development of predictive model for online review credibility. To build the predictive model, we tried different classifiers including, Deep Leaning, Support Vector Machine, Decision Tree, K Nearest Neighbors and Logistic Regression by considering review extremity, review length, external consistency, review subjectivity, internal consistency and review readability as their inputs. The predictive model based on deep learning could reach the AUC score of 0.82, which showed a better performance compared to other classification models in our research. The developed model has been tested against two datasets: the first dataset was labelled based on Yelp filtering algorithm and the second one was labelled based on the crowd's perception towards the credibility of online reviews. The predictive model shows a good performance in both datasets, which highlights the importance and validity of the attributes used in this research to evaluate the credibility of online consumer reviews. To the best of our knowledge this is the first study to investigate the credibility of online reviews by considering consumers' perspective in a predictive model (Abedin et al. 2020).

The finding from this study shows that to develop a predictive model for the credibility of online reviews, we need to consider a variety of attributes. It means, no attribute alone can signify the credibility of an online review, and instead, several important attributes together (e.g., review external and internal consistency, review length, review readability, review extremity, and review objectivity) should be used to predict the credibility of a given review. To discuss this further, we use Figure 2 and provide three examples of X, Y and Z in this figure. As can be seen in example X (which is marked in Figure 2), when a review is objective, but it is an extreme review, inconsistent and has a medium length, the review does not consider to be credible, which shows that objectivity alone cannot signify the credibility of an online review. In addition, as shown in example Y - Figure 2, when a review has a long length and is among non-extreme ones, but it is subjective and inconsistent with other reviews, it is not found to be credible, which shows that considering length and extremity without other attributes cannot signify the credibility of a given review. In the example Z - Figure 2, when a review is consistent with other reviews and is among non-extreme ones but has a medium length and is written in a subjective language, is not considered to be credible. This suggests that consistency and extremity cannot signify the credibility of a given review. These findings suggest that users need to make decisions cautiously based on online reviews information – for instance, by reading different reviews to realize the convergence among a series of views towards a product or service. In the same vein, this study recommends consumers consider a variety of different attributes (e.g., the internal and external consistency, objectivity, length, readability and extremity) at the same time to make their decisions rather than only relying on one piece of information or just considering one criterion like the stars rating of an online review. This highlights that no single attribute alone can signify the credibility of online reviews, and in turn, credibility is a complex factor that a combination of attributes together should be used for its measurement.

The findings from this research make some important contributions. First, this study contributes to the state-of-the-art literature around the credibility of online reviews by incorporating consumers' perspective through 1) using the attributes they use while assessing the credibility of online reviews and 2) using the crowd's perception for labelling the second dataset. Second, the predictive model can help consumers, businesses, and e-commerce platforms to evaluate the credibility of their information and filter out fake reviews from their websites. This, especially, is more essential for businesses and platforms that do not have any tools or models to monitor the credibility of their online reviews, and in particular, do not capture information about their users. Evaluating the credibility of information is an important task for businesses and online review platforms because providing consumers with credible information could increase consumers' loyalty, trust and overall satisfaction (Wu et al. 2020).

As with any other research, this study has some limitations. First, the datasets used in this study to train the classifier and test its performance were relatively small. Second, we only focused on Yelp reviews because this dataset had labels and therefore it was possible for us to develop the predictive model. Thus, the findings of this research and its generalizability should be treated with caution. It is highly recommended that future research considers other online review platforms (e.g., TripAdvisor, Amazon, etc.), representing different domains to expand the generalizability of the results of this research, which will require getting labels for this data to validate the results. For the future work, we plan to extend this study by conducting a series of ANOVA (analysis of variance) to investigate the different characteristics of fake and credible reviews in both datasets to explore their similarities and differences.

6 References

Abedin, E., Mendoza, A., and Karunasekera, S. 2019a. "Towards a Credibility Analysis Model for Online Reviews.," (*PACIS 2019*).

Abedin, E., Mendoza, A., and Karunasekera, S. 2019b. "What Makes a Review Credible? Heuristic and Systematic Factors for the Credibility of Online Reviews.," (*ACIS 2019*).

Abedin, E., Mendoza, A., and Karunasekera, S. 2020. "Credible Vs Fake: A Literature Review on Differentiating Online Reviews Based on Credibility,"*(ICIS 2020)*.

Abedin, E., Mendoza, A., and Karunasekera, S. 2021. "Exploring the Moderating Role of Readers' Perspective in Evaluations of Online Consumer Reviews," *Journal of Theoretical and Applied Electronic Commerce Research* (16:7), pp. 3406-3424.

Aghakhani, N., Oh, O., Gregg, D. G., and Karimi, J. 2020. "Online Review Consistency Matters: An Elaboration Likelihood Model Perspective," *Information Systems Frontiers*), pp. 1-15.

Ansari, S., and Gupta, S. 2021. "Customer Perception of the Deceptiveness of Online Product Reviews: A Speech Act Theory Perspective," *International Journal of Information Management*).

Barbado, R., Araque, O., and Iglesias, C. A. 2019. "A Framework for Fake Review Detection in Online Consumer Electronics Retailers," *Information Processing & Management*).

Basheer, I. A., and Hajmeer, M. 2000. "Artificial Neural Networks: Fundamentals, Computing, Design, and Application," *Journal of microbiological methods* (43:1), pp. 3-31.

Bischof, H., Schneider, W., and Pinz, A. J. 1992. "Multispectral Classification of Landsat-Images Using Neural Networks," *IEEE transactions on Geoscience and Remote Sensing* (30:3), pp. 482-490.

Cheung, C. M.-Y., Sia, C.-L., and Kuan, K. K. 2012. "Is This Review Believable? A Study of Factors Affecting the Credibility of Online Consumer Reviews from an Elm Perspective," *Journal of the Association for Information Systems* (13:8), p. 618.

Dong, L.-y., Ji, S.-j., Zhang, C.-j., Zhang, Q., Chiu, D. W., Qiu, L.-q., and Li, D. 2018. "An Unsupervised Topic-Sentiment Joint Probabilistic Model for Detecting Deceptive Reviews," *Expert Systems with Applications* (114), pp. 210-223.

Eslami, S. P., Ghasemaghaei, M., and Hassanein, K. 2018. "Which Online Reviews Do Consumers Find Most Helpful? A Multi-Method Investigation," *Decision Support Systems* (113), pp. 32-42.

Filieri, R. 2016. "What Makes an Online Consumer Review Trustworthy?" *Annals of Tourism Research*).

Fitzpatrick, E., Bachenko, J., and Fornaciari, T. 2015. "Automatic Detection of Verbal Deception," *Synthesis Lectures on Human Language Technologies* (8:3), pp. 1-119.

Goodfellow, I., Bengio, Y., and Courville, A. 2017. "Deep Learning (Adaptive Computation and Machine Learning Series)," *Cambridge Massachusetts*), pp. 321-359.

Heaton, J. 2008. *Introduction to Neural Networks with Java*. Heaton Research, Inc.

Hodge, D. R., and Gillespie, D. F. 2007. "Phrase Completion Scales: A Better Measurement Approach Than Likert Scales?," *Journal of Social Service Research* (33:4), pp. 1-12.

Hu, N., Bose, I., Gao, Y., and Liu, L. 2011. "Manipulation in Digital Word-of-Mouth: A Reality Check for Book Reviews," *Decision Support Systems* (50:3), pp. 627-635.

Jindal, N., and Liu, B. 2007. "Analyzing and Detecting Review Spam," *Seventh IEEE International Conference on Data Mining (ICDM 2007)*: IEEE, pp. 547-552.

Jindal, N., and Liu, B. 2008. "Opinion Spam and Analysis," *Proceedings of the 2008 international conference on web search and data mining*: ACM, pp. 219-230.

Krimpenis, A., Benardos, P., Vosniakos, G.-C., and Koukouvitaki, A. 2006. "Simulation-Based Selection of Optimum Pressure Die-Casting Process Parameters Using Neural Nets and Genetic Algorithms," *The International Journal of Advanced Manufacturing Technology*).

Kumar, N., Venugopal, D., Qiu, L., and Kumar, S. 2018. "Detecting Review Manipulation on Online Platforms with Hierarchical Supervised Learning," *Journal of Management Information Systems* (35:1), pp. 350-380.

LeCun, Y., Bengio, Y., and Hinton, G. 2015. "Deep Learning," *nature* (521:7553), pp. 436-444.

Leung, S.-O. 2011. "A Comparison of Psychometric Properties and Normality in 4-, 5-, 6-, and 11-Point Likert Scales," *Journal of social service research* (37:4), pp. 412-421.

Liu, Y., Pang, B., and Wang, X. 2019. "Opinion Spam Detection by Incorporating Multimodal Embedded Representation into a Probabilistic Review Graph," *Neurocomputing* (366), pp. 276-283.

Loria, S. 2018. "Textblob Documentation," *Release 0.15* (2), p. 269.

Lu, Y., Zhang, L., Xiao, Y., and Li, Y. 2013. "Simultaneously Detecting Fake Reviews and Review Spammers Using Factor Graph Model," *Proceedings of the 5th annual ACM web science conference*, pp. 225-233.

Luca, M., and Zervas, G. 2016. "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud," *Management Science* (62:12), pp. 3412-3427.

Luo, C., Luo, X. R., Xu, Y., Warkentin, M., and Sia, C. L. 2015. "Examining the Moderating Role of Sense of Membership in Online Review Evaluations," *Information & Management*).

Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., and Ghosh, R. 2013a. "Spotting Opinion Spammers Using Behavioral Footprints," *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 632-640.

Mukherjee, A., Venkataraman, V., Liu, B., and Glance, N. 2013b. "What Yelp Fake Review Filter Might Be Doing?," *Seventh international AAAI conference on weblogs and social media*.

Norton, M., and Uryasev, S. 2019. "Maximization of Auc and Buffered Auc in Binary Classification," *Mathematical Programming* (174:1), pp. 575-612.

Ott, M., Cardie, C., and Hancock, J. T. 2013. "Negative Deceptive Opinion Spam," *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies*, pp. 497-501.

Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. 2011. "Finding Deceptive Opinion Spam by Any Stretch of the Imagination," *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*: Association for Computational Linguistics, pp. 309-319.

Plotkina, D., Munzel, A., and Pallud, J. 2020. "Illusions of Truth—Experimental Insights into Human and Algorithmic Detections of Fake Online Reviews," *Journal of Business Research*).

Rayana, S., and Akoglu, L. 2015. "Collective Opinion Spam Detection: Bridging Review Networks and Metadata," *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*, pp. 985-994.

Rout, J. K., Dalmia, A., Choo, K.-K. R., Bakshi, S., and Jena, S. K. 2017a. "Revisiting Semi-Supervised Learning for Online Deceptive Review Detection," *IEEE Access* (5), pp. 1319-1327.

Rout, J. K., Singh, S., Jena, S. K., and Bakshi, S. 2017b. "Deceptive Review Detection Using Labeled and Unlabeled Data," *Multimedia Tools and Applications* (76:3), pp. 3187-3211.

Salminen, J., Kandpal, C., Kamel, A. M., Jung, S.-g., and Jansen, B. J. 2022. "Creating and Detecting Fake Reviews of Online Products," *Journal of Retailing and Consumer Services* (64), p. 102771.

Senter, R., and Smith, E. A. 1967. "Automated Readability Index," CINCINNATI UNIV OH.

Spooren, P., Mortelmans, D., and Denekens, J. 2007. "Student Evaluation of Teaching Quality in Higher Education: Development of an Instrument Based on 10 Likert-Scales," *Assessment & Evaluation in Higher Education* (32:6), pp. 667-679.

Weise, K. 2011. "A Lie Detector Test for Online Reviewers," *Bloomberg Business Week*).

Wu, H., and Leung, S.-O. 2017. "Can Likert Scales Be Treated as Interval Scales?—a Simulation Study," *Journal of Social Service Research* (43:4), pp. 527-532.

Wu, Y., Ngai, E. W., Wu, P., and Wu, C. 2020. "Fake Online Reviews: Literature Review, Synthesis, and Directions for Future Research," *Decision Support Systems* (132), p. 113280.

Zhang, D., Zhou, L., Kehoe, J. L., and Kilic, I. Y. 2016. "What Online Reviewer Behaviors Really Matter? Effects of Verbal and Nonverbal Behaviors on Detection of Fake Online Reviews," *Journal of Management Information Systems* (33:2), pp. 456-481.

Zhang, L., Wu, Z., and Cao, J. 2017. "Detecting Spammer Groups from Product Reviews: A Partially Supervised Learning Model," *IEEE Access* (6), pp. 2559-2568.

Copyright