information

Article

Semantic Features-Based Discourse Analysis Using Deceptive and Real Text Reviews

Husam M. Alawadh [1], Amerah Alabrah [2], Talha Meraj [3,*] and Hafiz Tayyab Rauf [4]

[1] Department of English Language and Translation, College of Languages and Translation, King Saud University, Riyadh 11451, Saudi Arabia
[2] Department of Information Systems, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia
[3] Department of Computer Science, COMSATS University Islamabad—Wah Campus, Wah Cantt 47040, Pakistan
[4] Centre for Smart Systems, AI and Cybersecurity, Staffordshire University, Stoke-on-Trent ST4 2DE, UK
* Correspondence: talhameraj32@gmail.com

Abstract: Social media usage for news, feedback on services, and even shopping is increasing. Hotel services, food cleanliness and staff behavior are also discussed online. Hotels are reviewed by the public via comments on their websites and social media accounts. This assists potential customers before they book the services of a hotel, but it also creates an opportunity for abuse. Scammers leave deceptive reviews regarding services they never received, or inject fake promotions or fake feedback to lower the ranking of competitors. These malicious attacks will only increase in the future and will become a serious problem not only for merchants but also for hotel customers. To rectify the problem, many artificial intelligence–based studies have performed discourse analysis on reviews to validate their genuineness. However, it is still a challenge to find a precise, robust, and deployable automated solution to perform discourse analysis. A credibility check via discourse analysis would help create a safer social media environment. The proposed study is conducted to perform discourse analysis on fake and real reviews automatically. It uses a dataset of real hotel reviews, containing both positive and negative reviews. Under investigation is the hypothesis that strong, fact-based, realistic words are used in truthful reviews, whereas deceptive reviews lack coherent, structural context. Therefore, frequency weight–based and semantically aware features were used in the proposed study, and a comparative analysis was performed. The semantically aware features have shown strength against the current study hypothesis. Further, holdout and k-fold methods were applied for validation of the proposed methods. The final results indicate that semantically aware features inspire more confidence to detect deception in text.

Keywords: credibility check; discourse analysis; frequency features; semantically aware features

check for updates

Citation: Alawadh, H.M.; Alabrah, A.; Meraj, T.; Rauf, H.T. Semantic Features-Based Discourse Analysis Using Deceptive and Real Text Reviews. *Information* **2023**, *14*, 34. https://doi.org/10.3390/info14010034

Academic Editor: Willy Susilo

Received: 11 November 2022
Revised: 10 December 2022
Accepted: 22 December 2022
Published: 6 January 2023

1. Introduction

Social media has become a quick source of information. Users rely on an internet network–based perception of online consumers [1]. The ease of use of social media inspires confidence in the information given on platforms such as Facebook and Twitter [2]. However, the fake information spreading on these platforms is a rising phenomenon resulting in misguiding users. This phenomenon has been coopted by individuals and even organizations that plant unrealistic promotion and criticism in reviews. This creates an environment of uncertainty around internet-based hotel reservations and many other e-commerce websites. Many consumers find online reviews credible when choosing a hotel [3] because authentic reviewers freely and openly share and discuss their experiences regarding certain hotel services.

The rapid spread of fake news and reviews has become a big challenge due to the abundance of information on the internet [4]. Manual detection of deception in online reviews is evidently impractical due the labor intensive and time consuming nature of analyses associated with it. Therefore, significant effort needs to go into developing efficient credibility checks for reviews and news [5]. Potential customers on online hotel reservation websites use people's reviews to make their decisions. Consumer feedback on hotels and other hospitality outlets accounts enable businesses to become aware of the quality of their services and make improvements when needed [6]. This open feedback creates a narrative, a description of services used by customers [7].

Reviews on websites such as Airbnb and Amazon share users' feelings to describe their experiences. This open sharing of feedback ultimately presents an opportunity for analysis by the specialists from different fields including discourse analysis. Discourse specialists perform discourse analysis to obtain the social context of the reviews. It narrates people's opinions and interprets implicit and concealed meanings [8]. This open sharing helps service providers, but it also creates an opportunity for fake reviewers to deceive people, which generates uncertainty around online portals [9]. The fake reviewers spread fake positive reviews about products of a merchant or damage the reputation of a competitor's product. Both real and deceptive reviews contain complex clues to their authenticity [10]. The real reviews contain people's feelings, expressed in genuine terms. They typically reflect where the reviewer visited and which services they used from a particular hotel [7]. These reviews which can be positive or negative/critical reviews are often encouraged by merchants, as they provide suggestions for improvement.

Fake or deceptive reviews do not express people's feelings; the verbal leakage in them [11] represented in the absence of feelings, contextual discontinuity, and a lack of coherence indicate their falsehood. Discourse analysis helps to decipher meaning and sentiment beyond the sentence level [12]. Therefore, to perform credibility checks on reviews, a foolproof, trustworthy automated tool is essential to check the truthfulness of a given text [13].

Online Fake promotion of products and services compromises the process of experience sharing thereby discrediting the genuine reviews and feedback. Open sharing of consumer feedback is beneficial, but it can also become very risky. It could highlight the unique attributes of a certain hotel, both positive and negative [14].

Not all hotels and restaurants invite open sharing; some offer a multiple-choice questionnaire to obtain people's feedback. These questionnaires do not give customers the opportunity to analyze services, making the feedback itself questionable. Therefore, open sharing is encouraged by the discourse community as well [8]. It helps distinguish between real and fake reviews.

However, finding the hidden feelings of visitors is a very complex task that is made more difficult by fake reviews. Many sentiments analysis tools for performing discourse analysis have been conducted, but the capabilities of artificial intelligence are improving day by day, discovering robust algorithms The fake reviewers have a limited vocabulary, whereas real reviews have more variety in terms of feelings, words, and description. Similarly, a lack of context and incoherent sentences help to identify fake content. The frequency of certain words in reviews could be used to detect fraudulent reviews. A conflict between real and fake services highlighted in reviews is also an indicator of illegitimacy [15].

However, while these linguistic cues can identified and classified, such process, if done manually, is time consuming and tedious and by the time they are identified, the negative impact of fake reviews may take place. An automated solution is needed that could identify fake and real reviews using multiple features in each document. To encourage the discourse analysis, to remove the un-certainty upon deceptive texts given on hotel reviews, the proposed study utilized two different types of methods. The hidden and in-depth meaning is detected with the semantic meaning in it with use of deep features.

The current study has used two different approaches to identify fake and real reviews. The dataset used by the current study contains reviews of different places by different people. It is well balanced. However, a comparative machine-learning approach is used that includes a word-frequency feature and a new, deep-learning-based standard encoding method of text embedding. The current study contributed to the literature in the following ways:

- It highlighted the deep and semantic attributes of fake and real reviews to identify them.
- It performed a comparison of word frequency–based features and deep-learning features to understand the attributes of fake versus real reviews. It is concluded that semantic awareness in text is more important as compared to frequency count only.
- It used a credibility check on reviews that could be used as a discourse analysis tool.
- The current study used methods that are robust and could be applied on other e-commerce websites to check the validity of reviews.

This article is split into four sections: related work, methodology, results, and discussion, followed by a conclusion and future directions.

2. Related Work

In the hospitality sector, online presence and significance moved from being a trend into becoming a necessity. Most business owners and clients rely on this online presence in identifying the quality of services, users' past experience, and booking/visiting decisions. Researchers find the discourse shared on social media networks about hospitality services and particularly that of online reviews an important venue for research. In this section, we highlight some research studies which have looked into that. In a study conducted by [16], reviews from the TripAdviser website were collected and they were used to analyze customer behavior via T-LAB software. The main goal of this study is to measure the key indicator for the evaluation of hotel services. The data comprised 1,143,631 user reviews, which are extracted from 35,611 hotels in different cities in Brazil. The study reported that reviews typically indicate the location, quality of food, opening and closing hours, the variety of foods, etc., and that such information can be utilized to measure the quality of any restaurant.

To bridge a clear research gap evident in the lack of business needs analysis that is based on online hotel reviews, Perinotto et al. [17] applied a theoretical framework to analyze customer decisions towards the online purchase of hotel reservation are influenced by online reviews. The unified concept of price and social influence towards positive and negative thoughts are applied with the addition of trust and perceived risk assessments. A questionnaire is used to collect data from 195 different residents of Ceará, Brazil. This research indicates that important factors towards online hotels' reservation, such as price and social influence, are the most important factors.

Similarly, to analyze the customer behavior on the Airbnb website, the manual collection of user reviews is performed using NVivo software [18]. It includes 2353 reviews extracted from different 506 lodging offered in 2019–2020. The final results demonstrated sentiments of people towards different services such as those people who have given negative thoughts have more critical analysis, which is helpful in order to improve hotel services. It also indicates that when data are evaluated based upon other factors such as gender, location, offer types, and host types, they demonstrated different sentiments.

The studies discussed above demonstrate that online hotel reservation systems have different business needs in order to improve their services. However, to analyze that user behavior and sentiment toward a hotel service, its public data are used. This public data or provided feedback on hotel websites includes positive and negative thoughts. These online provided feedbacks are being accessed by fraudsters of different competitors to provide a fake promotion and to degrade hotels.

Therefore, detection of fake and deceptive reviews has also been an ongoing research endeavor for many years. The techniques used to detect fake or deceptive reviews can be classified in three main categories: user-centric, feedback-centric, and network-centric approaches. In a user-centric approach, reviewer profiles are checked, including factors such as how authentic they are, what ratings they have given each service, and how popular they are. Reviewers' feedback regarding services is taken into consideration, as it contains a variety of feelings in each review, whereas fake reviews lack emotional content—this is checked using linguistic cues. A network-centric approach is different; it examines links between the objects involved in reviewing services and attempts to detect conflict between reviews by fake and real users.

Few of the recent studies discussed here have highlighted the gaps between previous studies and the challenges that need to be addressed until now. Another study used a credibility score which is assigned using a section of text [5]. This text is first given to search for the relevant documents on it from the web. The interaction, language style, and trustworthiness of the text based on its claimed stance are checked against other sources. A credibility score is calculated using a predetermined lexicon. It demonstrated various combination-based accuracy scores for finding false and true claims in reviews. It outperformed the macroaverage accuracies on the pipeline method as 82%, CRF method as 80%, and LSTM as 78.09% . David F. Larcker et al. [19] reported that there are significant differences between deceptive and real calls. The positive word usage in CEO calls and non-CEO calls is different. Similarly, deceptive calls contained different facts about a service from real calls. This approach uses word frequency as a feature of real and deceptive language identification by way of linguistic cues and financial variables. The summarized recent development using different methodologies and their corresponding results are shown in Table 1.

Table 1. Summary of recently applied studies on deceptive text detection.

References	Year	Methods	Dataset	Results
[5]	2018	Credibility Score assessment via web-sources based upon user input	Web-Sources and User inputs of fake and true claims	Macro-avg accuracy Pipeline = 82%
[5]	2018	Credibility Score assessment via web-sources based upon user input	Web-Sources and User inputs of fake and true claims	Macro-avg accuracy CRF = 80%
[5]	2018	Credibility Score assessment via web-sources based upon user input	Web-Sources and User inputs of fake and true claims	Macro-avg accuracy LSTM = 78.09%
[20]	2019	Different Features TF-IDF, user-centric and bag of words with ML classifiers	Yelp dataset via web-scrapping	Highest achieved F1-score Ada-Boost = 84%
[21]	2021	Word, sentence and chunk level model	Different public datasets	-
[22]	2022	Chinese language model via dynamic features	Samsung case reviews and restaurant reviews datasets	Precision = 0.92, Recall = 0.91, F-score = 0.91
[16]	2022	TOURQUAL Protocol visa T-LAB software is applied to find hotel quality control indicators	1,143,631 user reviews extracted from TripAdvisor website	
[17]	2022	A conceptual framework applied based upon UTAUT2 model	195 residents data collected using digital questioner	
[18]	2022	Airbnb customer behavior analyzed using Nvivo software	2353 reviews collected via manual coding	

Another study [23] claims that the fake reviewers used different writing styles to promote or disparage popular service providers. Therefore, it is necessary to use style-based features for reviews. The verbal and syntactical loss between real and fake reviews is identified using optimized sequential minimal function via the support vector machine (SVM). It achieved 84% F-measure on a combination of lexical and syntactical features, while the naïve Bayesian method showed a 74% F-measure. Conflicts appear when services are discussed that are not actually provided by the hotel being reviewed [15]. This conflict could be used to detect a fake review.

On some websites, user profiles are maintained, along with different badges such as “most valuable customer", indicating more interaction and/or engagement on the website. These profiles are rated by other consumers as well. Therefore, their reviews are mostly helpful and critical ones; due to these users’ popularity, people believe their reviews are honest [24]. However, these could be planted reviews. They can be detected by their wording and promotion of fake services along with real services. They can be hard to identify manually; therefore, an automated detection method is needed.

Network-based fraud detection is another way to perform a credibility check. The neighborhood is likely to be used in image-processing concepts as well because it provides more appropriate values for an object under consideration. A 2-hop network is used to identify reviewer trustworthiness by social network–based credibility checks [25]. Data-scraping tools are used to collect people’s reviews from four different cities, and a feature framework is designed that is later used in the classification of fake reviews [20].

In a feature framework, the bag of words, TF-IDF , and user-centric features are used. A combination can also be applied. For the sake of classification, the random forest, decision tree, logistic regression, gaussian kernel-based naïve Bayesian method, and the Ada-boost classifiers are used, where Ada-boost shows the highest F-score of 82%. Single-dimension features for fake review detection are discouraged, whereas a 3D feature selection approach using information gain, chi-square tests, and XG-boost methods is successfully applied [26]. The deception score, a user’s personal behavior, and the text features are used in this study. The evaluation measure, such as AUC, macroaverage, or WF score, are reported at the end. Two Amazon review datasets are used in this study.

Deep learning–based, multifeature extraction and fusion are performed on three different unbalanced datasets [27]. The deep learning–based models used in this study are textCNN, GRU, and attention-based local and temporal; weighted semantic features are also used. A new feature fusion strategy was adopted for feature selection. Accuracy, precision, recall, F-score, and AUC were reported for balanced and unbalanced datasets on all three types of datasets used in this study. The fusion strategy demonstrated comparatively better results on balanced as compared to unbalanced datasets.

In another study, Chinese language–based deceptive reviews were collected, and dynamic features from these reviews were extracted [22]. There are claims that the fake reviews are misleading customers to purchase online goods that do not match the specifications described in those reviews. However, the pre-trained Chinese-language model achieved 0.92 (precision), 0.91 (recall), and 0.91 F-score measures. The results were 20% higher than those of previous studies, showing the effectiveness and robustness of the methods employed in the current study.

Another study [28] uses various product parameters as features to identify deceptive reviews. Text length, nature of the text, rating on the review, verified purchase status, use of pronouns in text, and other product types were taken into consideration. SVM and naïve Bayesian classifiers were used. The accuracy rate showed satisfactory performance by the current study. The authors used chunks of text and emotions on a sentence level [21]. Sentence, body, and emotion levels were incorporated in classification. The experimental results outperformed utilized datasets on previous studies.

We have discussed older and more recent studies on deceptive review detection that adopted the three types of approaches discussed above. Recent studies mainly utilize the deep learning–based methods, which are also encouraging for the detection of deceptive reviews, as they uncover the hidden sentiments in both fake and real reviews.

The previous studies demonstrated the importance of frequency-based features that showed the fraudster’s immaturity. Many other factors, such as verbal loss, contextual loss, and coherence loss, could be used as linguistic cues in order to distinguish the deceptive versus real reviews. The frequency-based features and deep learning–based, review-centric features are used in the current study, and a comparative analysis of both approaches is performed below.

3. Materials and Methods

The recent need for deceptive review detection has attracted the attention of scholars worldwide. Many of the recent developments on fake review detection employed semantics, word frequency, user profiles, and network-based-deep-learning methods. However, in the framework of the current study, two different approaches are applied for a comparative analysis of deceptive review detection methods. The primary steps are shown in Figure 1.

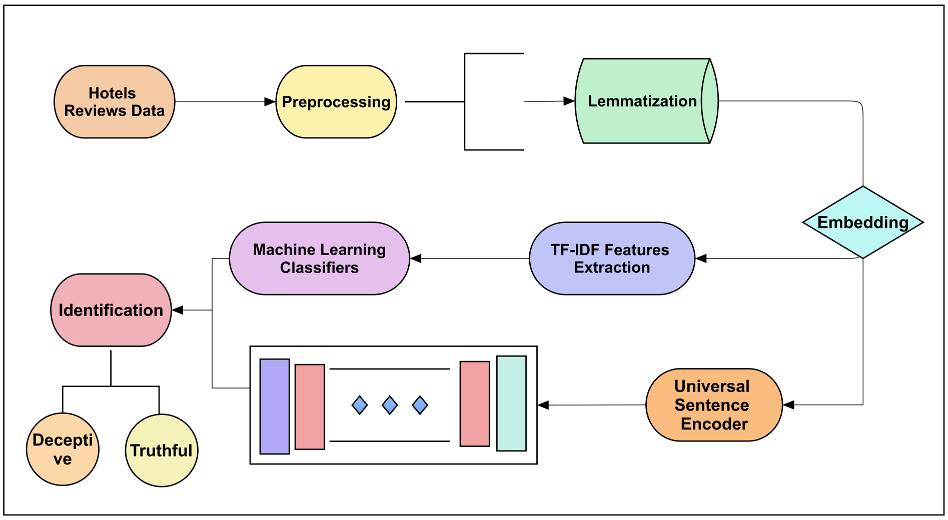

Figure 1. Flowchart containing primary steps applied in proposed framework.

The hotel reviews are firstly processed into the pre-processing phase by removing URLs, the lower case of words, removal of ambiguous words or repetitive words which are usually used when people talk about something positively and negatively. Therefore, the lemmatization after pre-processing step is applied. This refined text is then fed to the embedding system to encode the frequency-based and semantic feature-based two methods.

In method-1, the frequency-based TF-IDF method of feature extraction is adopted, and these features are given to machine-learning classifiers. In method-2, the universal standard encoding (USE) layer has been used, and then dropout and soft-max activation layers are employed to identify the deceptive and real reviews.

The TF-IDF features containing word count-based features are fed to ML classifiers with various splits of data. However, in method-2, the deep learning layers are used and softmax layer is used in the classification of real and fake reviews. For the sake of more validity, the holdout and K-Fold validation schemes are applied.

3.1. Preprocessing and Lemmatization

The data are processed into a final, refined form that excludes words that are not English. Expression-based string refining excluded capitalized and small words.

Using the refined English text, the capitalized words are converted into lowercase letters, as that makes the words equivalent without changing the original text. In a natural context, when we type or speak, we include criticism, appreciation, and other context. Moreover, we sometimes express the same ideas using multiple sentences of similarly weighted words, resulting in unwanted redundancy. Therefore, we need to lemmatize similar words to reduce redundant information.

To remove the redundancy from expressions, lemmatization is performed, and stop words are also removed [29]. The Natural Language Toolkit (NLTK) is used for lemmatization, along with stop words Application Programming Interface (API) . Stemming is also used to remove similar words, but lemmatization is preferred over stemming as it uses the structural and contextual word replication of the given text. Lemmatization makes the document more meaningful, as it is a subset of WordNet, which contains millions of English words. This method recognizes only those words found in the WordNet library. However, it mainly recognizes the given text and refines it to obtain the most appropriate credibility score.

3.2. Term Frequency–Inverse Document Frequency Features

In related work, we have observed that the frequency of words in real and fake text matters a lot. Deceptive reviews may not contain strong, factual praise of any services, whereas real reviews contain actual facts. Therefore, the words describing such facts and the users' sentiments are significant. To utilize this word significance in the current study, the frequency of words is examined using a TF-IDF-based feature sparsity matrix. It contains the weighted score of each word occurring in the document and, based upon that, a sparsity matrix. The final feature vector size is X*5510, where X is the training and testing instance size. The sparsity matrix contains the word score as calculated by TF-IDF. The method involves the following steps.

$$d_f(t) = f_t(docs) \tag{1}$$

Equation (1) explains how the term frequency works to document words. The term frequency is similar to a word's occurrence in the document, but not identical, in the case of $d_f(t)$. It is the frequency of a word occurring in all documents that is found at least once in a single document. Therefore, on the right side of Equation (1), the frequency of term $t(f_t)$ is found in documents $(docs)$.

$$idf(t) = log(\frac{N}{d_f(t) + 1}) \tag{2}$$

Equation (2) explains how the inverse document frequency, or the weighting formula for each term, is calculated. N represents the body length divided by the frequency of a certain item. It is the inverse of any value occurring too many times, such as stop words or "is," "am," "are"—words found many times that may disturb the occurrence of rarely found terms in the document. It lowers the weight of words that are found too many times. Unique or meaningful words thus have more weight as compared to other words. The finalized version of the full method TF-IDF is found in Equation (3).

$$tf - idf = d_f(t) * idf_t \tag{3}$$

Equation (3) represents the whole formula for the frequency of term t and its corresponding weight as calculated using Equation (2). The weight and frequency of a term based on a sparsity matrix for each document plays the role of a feature matrix of training and testing instances. In this way, important and unique facts and realistic and fake feelings could be given corresponding weights if the fake reviews contain less confident words or feelings. The terms of a fake review will have less weight, whereas more meaningful facts and feelings based on strong arguments will be appropriately weighted by occurrence. In this way, the TF-IDF method features could give us credible wording-based scores.

3.3. Machine-Learning Classification

The sparsity matrix–based input is from known machine-learning classifiers. We used classifiers for different domains to obtain more reliable results. In the current study, the naïve Bayesian probability-based classifier, the SVM radial basis function kernel, logistic regression-based analysis, and decision tree and random forest methods have been applied. In the naïve Bayesian method, the sparsity matrix is passed as it is, where in all other methods, the input was fed as a standardized scalar 1D array by converting the sparsity matrix into a single input–instance matrix. However, the performance of naïve Bayesian remains the highest, which could be due to the sparsity matrix input. The results are discussed in Section 4.

3.4. Universal Sentence Encoding (USE) Method

Many text-to-numeric encoding methods are used in computer-assisted tasks. Computers can understand text in a numeric or binary form; the text-to-digitization method is simply known as embedding, where one hot encoding is the simplest method of representing text in a numeric form. However, the text could be represented in a numeric format instead of digitization when we use any other semi-supervised method, which could make the semantic classification and credibility check more efficient.

Similarly, a USE method is introduced that not only converts the text into numeric form but also gives the sentence a semantic similarity score using a pretrained deep averaging network (DAN).The visual representation to obtain the similarity index based upon the USE method is shown in Figure 2.

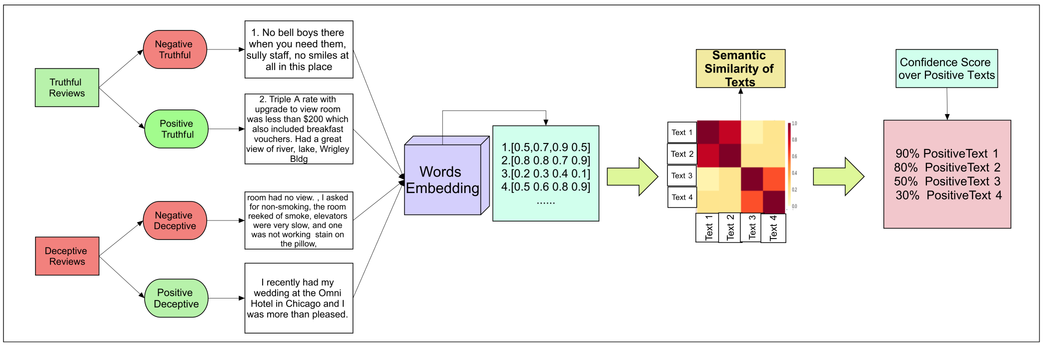
Truthful Reviews
Negative Truthful
Positive Truthful
Deceptive Reviews
Negative Deceptive
Positive Deceptive
1. No bell boys there when you need them, sully staff, no smiles at all in this place
2. Triple A rate with upgrade to view room was less than $200 which also included breakfast vouchers. Had a great view of river, lake, Wrigley Bldg
room had no view. , I asked for non-smoking, the room reeked of smoke, elevators were very slow, and one was not working stain on the pillow,
I recently had my wedding at the Omni Hotel in Chicago and I was more than pleased.
Words Embedding
1.[0.5,0.7,0.9 0.5]
2.[0.8 0.8 0.7 0.9]
3.[0.2 0.3 0.4 0.1]
4.[0.5 0.6 0.8 0.9]
......
Semantic Similarity of Texts
Text 1
Text 2
Text 3
Text 4
1.0
0.8
0.6
0.4
0.2
0.0
Confidence Score over Positive Texts
90% PositiveText 1
80% PositiveText 2
50% PositiveText 3
30% PositiveText 4

Figure 2. Working Scenerio of USE method in order to calculate similarity co-efficient.

It is trained on a benchmark of STS that contains words and sentences collected from news, forums, and captions. It is quite a complex task and gives us a standard, equalized embedding of a given English text input. As it is trained on big data that contains sentence similarity scores, DAN is far better than a simple text-to-numeric representation. It gives us the semantic meaning of any sentence as well, which leads to a clearer understanding of the text, such as whether it is a question or a positive or negative review and the criticism on any particular point.

Semantic similarity score–based text embedding could be used to classify the meaning of a text used in the current study to obtain both deceptive and real meanings using deep neural networks. The designed deep neural network layers and description are discussed below.

3.5. USE Layer–Based Deep-Learning Classification

In normal text embedding and text classification tasks, the embedding layer is used from Keras or TensorFlow libraries, where the embedding gives the numeric representation on which the normalization or activation function–based neural network is designed. In the current study, rather than applying a simple embedding layer, the USE layer [30]—as described above—is chosen as the text embedding input that returns the semantically aware input to the network, and then a simple feed-forward network is applied where the fine-tuned proposed network is shown in Table 2.

Table 2. USE-Based proposed deep neural network.

Serial Number	Layer Name	Input Parameter	Activation Function
1	Universal Sentence Encoding	Text input	-
2	Dropout	0.2	-
3	Dense	128	ReLU
4	Dense	64	ReLU
5	Dense	32	ReLU
6	Dense	16	ReLU
7	Dense	1	Sigmoid

Table 2 shows a simple feed-forward neural network; the other layers, such as pooling and batch normalization, are less important than the USE layer, which gives us the important semantic scores that may be lost in batch normalization and pooling layer operations. The dense layer plays the role of a simple hidden layer, where a comparison of denser or weighted neurons to less dense neurons is performed, which could be called a top-to-bottom approach. Additionally, a ReLU layer of activation is applied, in which the negative values will be converted to zero. The ReLU function operational equation is shown in Equation (4).

$$ReLU = \begin{array}{lll} if x \leq 0 & then & x = 0 \\ if x > 0 & then & x = x \end{array} \tag{4}$$

In Equation (4), x shows the input value of x, the output of any weight coming from the upper layer. The hidden network consists of 5 to 8 layers, and the input node ranges from 128 to 16. The final probabilistic layer with sigmoid activation function uses the calculated weights to divide the data into classes.

4. Results and Discussion

The framework of the current study is divided into two methods: one experiment is conducted using frequency-weighted features fed to machine-learning classifiers, and the second applied the semantically aware encoding of text and classification by a simple feed-forward neural network.

4.1. Data Description

The dataset used in the current study was balanced and collected from different hotels. It contains 1600 total instances, 800 deceptive and 800 truthful [31,32]. The breakdown between positive and negative instances is shown in Figure 3. The dataset is abundantly applied in previous studies, where it contains real reviews based on a number of hotels of Chicago. The dataset is properly labeled by experts and most importantly the class imbalance is not present at all in it, which is beneficial in order to validate the machine learning applied methods.

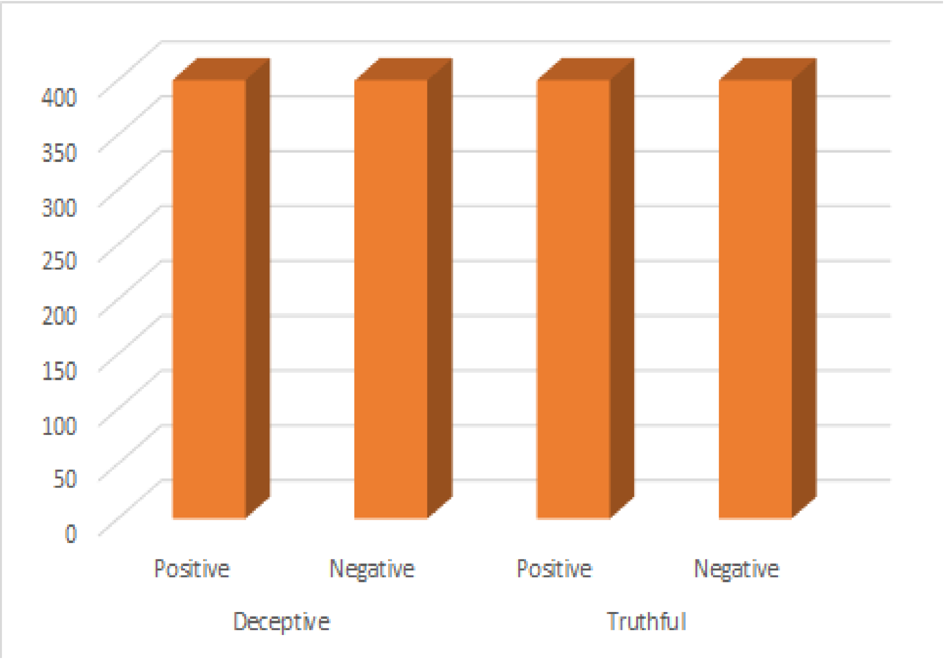

Figure 3. Class-wise distribution of dataset used in current study.

4.2. Evaluation Metrics

A discussion of evaluation measures used in our experiments will provide a better understanding of the current study's results. The related work, based on four evaluation metrics, was finalized for use in the current study. The evaluation measures are accuracy, precision, recall, and F-1 score. All measures are repeatedly used by authors in previous studies. Therefore, these evaluation metrics will provide a stronger comparison against results from previous studies. The mathematical representations of all metrics are shown in in Equations (5) through (8).

$$\text{Accuracy} = \left(\frac{TP + TN}{TP + FP + TN + FN}\right) \tag{5}$$

Accuracy is shown as a ratio of the sum of true positive and true negative instances over the sum of all positive, negative, true, and false instances.

$$\text{Precision} = \left(\frac{TP}{TP + FP}\right) \tag{6}$$

Precision is the true positive over positive instances detected truly and falsely. Similarly, recall is true positive over false negative and true positive summation.

$$\text{Recall} = \left(\frac{TP}{TP + FN}\right) \tag{7}$$

Recall is used to measure the true positive rate over the summation of true positive with false negative. It includes the affect of the true positive detection rate as compared to false negative instances.

$$\text{F1-Score} = 2 \times \left(\frac{recall \times precision}{recall + precision}\right) \tag{8}$$

Finally, the F-1 score contains the precision- and recall-based product over summation ratio with a multiplication of 2 as a constant. It is a more valid and appropriate measure of global performance, even in an unbalanced dataset. In the current study, the true positive instances are truthful reviews, and negative reviews are deceptive ones. It more importantly focuses on general fact of all true positive, negative and false positive, negative predicted instances with respect to their class balancing. Therefore, it is considered to be a more effective measure as compared to accuracy when the dataset exhibits a class imbalance problem.

4.3. Experiment 1

In the current study work, frequency-weighted words are used, and two validation techniques are applied. The holdout and cross validation techniques are applied on TF-IDF features. The results of both validation schemes are discussed below.

4.3.1. Holdout Validation

The validation techniques prove the experimentation tasks using appropriate evaluation measures. Our study uses two ways to prove its robustness and highlight any weaknesses. In the holdout method of validation, data are split into training and testing folds using manual approximation of each fold. The data are randomly divided into 70–30 (S-1), 60-40 (S-2), 50–50 (S-3), and 30–70 (S-4) splits, where the first number represents training data and the second represents testing data. The data splitting is performed using randomization to avoid any biases. The testing results are demonstrated on best-performed machine-learning classifiers.

As we can observe in Table 3, the performance of three classifiers remains slightly different in all splits. The naïve Bayesian method showed 86%, 86%, 87% and 84% accuracy measured in all splits of S-1, S-2, S-3 and S-4. The performance of each measures precision, recall, and F-1 score, which remain consistent. The recall also remains the same. Many important and decision-making evaluation measures the 85.5%, 85.5%, 86.5% and 83.5% F-1 score.

Table 3. Hold-out validation method results by applying TF-IDF features.

Sub-Exp.	Validation Split	Method	Acc. (%)	Prec. (%)	Rec. (%)	F1. (%)
S-1	70–30	Naïve Bayesian	86	86	86	85.5
		Linear Regression	81	84	81	80.5
		Random Forest	75	75	75	75
S-2	60–40	Naïve Bayesian	86	86	85.5	85.5
		Linear Regression	81	84	81.5	80.5
		Random Forest	74	74.5	74.5	74.5
S-3	50–50	Naïve Bayesian	87	87	87	86.5
		Linear Regression	81	83.5	81.5	80.5
		Random Forest	73	73.5	72.5	72.5
S-4	30–70	Naïve Bayesian	84	84	84	83.5
		Linear Regression	80	82.5	80.5	80
		Random Forest	70	71.5	70	69.5

The results consistently demonstrate that the naïve Bayesian method plays an important role in accurate classification, regardless of which split or how a model is trained on it. Second, the linear regression method is used as classification using a 0.5 threshold value for credibility. It showed values of 81%, 81%, 81% and 80%, which are again consistent regardless the amount of training data. The 84%, 84%, 83.5% and 82.5% scores showed in a similar manner the precision scores of the linear regression method. The recall values were 81%, 81.5%, 81.5% and 80.5%—again, similar values with slight differences. Random forest is a different category of classification; it shows 75%, 74%, 73% and 70% values for accuracy, which decreased when we used less training data.

The precision values were 75%, 74.5%, 73.5% and 71.55%; these values also decreased when we reduced the training data of random forest classifiers. Similarly, the random forest showed deviation while calculating its recall value. The last measure, F-1 score, showed distinctive results, as it combined the effects of precision and recall. The score in each experiment was reduced to 75%, 74.5%, 72.5% and 69.5%, showing an abrupt change in the performance of random forest classifiers when we reduced training data. The comparative analysis of the performance of these classifiers shows that a frequency-based method of checking the credibility of real reviews contains important information, and the naïve Bayesian method is the best classifier among those used in the current study, as it remains consistent in performance whether we reduce the training data or not.

4.3.2. K-Fold Validation

The second method of validation used in the current study is K-fold validation, where K represents a number greater than 2. In this method, the number-based folds are split for the whole data, and on each fold, the training and testing are performed in sequence. In this way, no data remains untrained and untested on a given classifier. The study makes four folds of whole data and shows the accuracy measure of each datum. Each fold is properly divided for training and testing; therefore, the accuracy measure is appropriate to show the performance of the classifier, as it does not contain any class imbalance.

Table 4 shows that the performance of naïve Bayesian does not remain highest when we applied this validation scheme. It can also be observed that the validation method can reduce the robustness of any classifier. This leads to a conclusion that in each conducted study, the multi-validation scheme should be applied. As we increased the folds of each classifier for training and testing, the performance of each classifier improved. In the case of naïve Bayesian, accuracy changed from 72% to 75.1% as the folds increased.

Table 4. K-Fold validation method results by applying TF-IDF features.

Methods	3-Fold	5-Fold	8-Fold	10-Fold
Naïve Bayesian	72%	72.8%	73.5%	75.1%
Linear Regression	83%	83.2%	82.9%	83.2%
Random Forest	73%	76.4%	78.3%	79.1%

The linear regression–based method of classification also showed deviation, but not as much, as it changes to 83%, 82.%, 82.9% and 83.2% values that are not incremental but it remains same.The worst-performing classifier in the holdout validation scheme was the worst here as well, but it is important to point that when we increased the folds of data, the accuracy increased from 73% to 79.1%, which is quite satisfactory for the performance of random forest classifiers. By looking into the performance of each classifier, we can say that if the training and testing of folds' increase, the performance of all classifiers would improve, as demonstrated in the results of the current study.

4.4. Experiment 2

The USE method of embedding was used in experiment 2. The feed-forward neural network was applied, and results are shown in Table 5. The various splits were applied on all data, and the results of less versus more training are shown.

Table 5. USE-based proposed deep neural network.

Sub-Experiment	Data Splits	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
S1	70–30	85.62	88.40	81.27	84.70
S1	60–40	82.96	83.05	81.16	82.10
S1	50–50	79.88	76.92	84.18	80.38
S1	30–70	79.29	78.54	79.96	79.24

The first split, S1, contains 70% training data and 30% testing data; accuracy was 85.62%, precision 88.40%, recall value 81.27%, and the F-1 score was 84.70%. All measures demonstrated satisfactory results on this split. The second split, S2, which included 60% training data and 40% testing data, demonstrated that less training led to less accuracy and precision. The recall value was almost equal to that of S1. The F-1 score showed 82.10% accuracy—less than the S1 split. S3 consisted of 50% training data and 50% testing data.

Accuracy was 79.88%, precision 76.92%, recall 84.18%, and the F-1 score was 80.38%. The recall value was an improvement over S1 and S2 scores. The F-1 score is more significant in that it shows a lower score by combining the effect of recall and precision. The S4 split contained 30% training data and 70% testing data. Surprisingly, the scores showed the robustness of the model; they were all still in the range of 80%. The F-1 score is, again, the most considerable measure, covering the class unbalancing issue as well as giving a global measure. The F-1 score is shown in Figure 4.

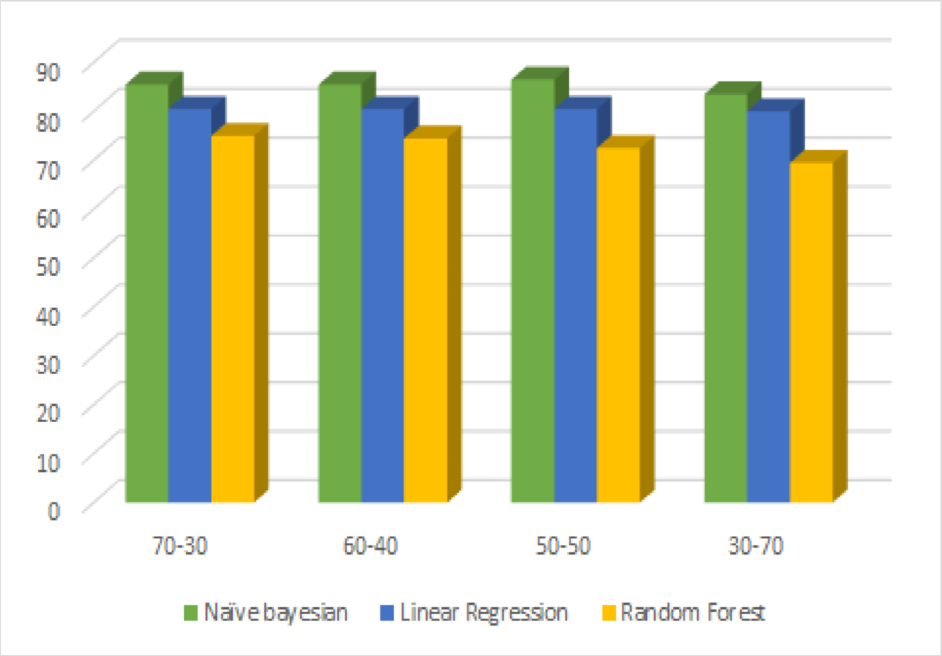

Figure 4. Visualization of F-1 score–based splits of three applied machine learning classifiers.

4.5. Comparison

We can observe in the comparison table that there are different kind of methods that have been applied on deceptive reviews' detection using the same dataset. The different evaluation metrics are also used where all comparative studies, as shown in Table 6, do not use as many metrics that are used by the proposed study such as accuracy, precision, recall and F1-score. The studies demonstrated a different measure-based performance measure of their approaches whereas the current study uses all metrics while demonstrating confidence in each aspect of evaluation.

Table 6. Comparison with state-of-the-art studies.

References	Method	Results (%)
[33]	Parts of speech, linguistic, word count and content-based features used	F1-Score = 83.7
[34]	Semi-supervised best-performed Naïve Bayesian	Accuracy = 85.21
[34]	Supervised best-performed Naïve Bayesian	Accuracy = 86.32

[35]	Bag of words feature selection method, best-performed Naïve Bayesian	Acc. = 87 Prec. = 52.78
		Rec. = 92.63

[36]	N-gram character-based deceptive review detection	25–75% split and F1. = 80%

Proposed Work	Semantic aware deep feature-based deceptive review detection	Acc. = 87 Prec. = 87
		Rec. = 87 F1. = 86.5

Compared to comparative studies, the F1-score in the first comparison is 83.7% whereas the current study shows a percentage of 86.5% in the same measure. The second comparative study uses supervised and semi supervised training, where the supervised training method demonstrated more accuracy of 86.32%. The proposed study demonstrated 87% score in accuracy, precision and recall which is higher than the third comparative study results as well.

The fourth comparison uses 25% of the data as training only and demonstrated 75% data testing results whereas the proposed study uses in its last split, the 30% for training and testing for 75% data where the F1-score remains 83.5%, a score that is higher than the comparative study. Therefore, the proposed study demonstrated more accurate and more authentic semantic aware feature-based robustness of deceptive reviews detection.

5. Theoretical Analysis

The related work shows that the frequency of words in the text is aligned with the truthfulness of a text, as the true reviews contain more facts and solid reasoning, whether criticism or praise. Therefore, a unique word weight–based credibility check could be performed on text reviews. Hypothesis-based TF-IDF features are extracted from deceptive and real reviews. The latter two validation techniques prove them true.

Deep-learning methods used in recent studies, as discussed above in the related work section, were employed. The four different splits and a universal semantically assisted layer were used with a simple feed-forward neural network, and the results were robust and trustworthy. This approach can be deployed in a real-time framework, as all amounts of the training data showed satisfactory results on the testing data. Therefore, semantically aware features are more useful, deployable, robust, and precise for all measures of accuracy.

6. Conclusions

As we become more reliant on the virtual space, fake reviews and news are constantly increasing both in quality and quantity. Identifying such deception has become a huge challenge in many fronts. Computer-automated methods of text recognition are used to identify spam texts, email, and other malicious attacks. The current study attempts to address this issue by using a balanced, real reviews dataset of hotel services comprising real as well as deceptive reviews. After text preprocessing, a comparative analysis of frequency-based features fed to machine-learning classifiers and semantically aware, deep embedding–based neural networks was performed to ascertain the credibility of the reviews. The two validation schemes applied in experiment 1 show the differences in performance when we reduced the training data. The cross-fold validation method in experiment 1 demonstrates that when we increase the folds, the performance of the classifier increases. In experiment 2, the performance of the deep-learning classifier is probably the same in terms of evaluation metrics, but it is more trustworthy as compared to TF-IDF features, as the classifiers using less training data showed better results when tested on even 70% of the data. Therefore, both experiments demonstrate that deep learning–based, semantically aware text features can be applied in real time on web portals to identify fraudulent reviews. One significant limitation of The current study is that it uses a limited reviews dataset against the deep learning approach. It is therefore suggested that future research uses a larger dataset to utilize the deep learning. Therefore, synthetic data generation and other methods could be used to increase the instances of the dataset. In our future research, we will attempt to address real time reviews deception, which means that we will develop methods to identify deceptive reviews as they are posted so that business owners are alerted about them and so they can act accordingly.

Author Contributions: Conceptualization, H.M.A.; Methodology, T.M.; Software, T.M.; Validation, A.A. and H.T.R.; Formal analysis, H.M.A.; Investigation, T.M.; Writing–review and editing, H.T.R.; Supervision, H.M.A.; Project administration, H.M.A. All authors have read and agreed to the published version of the manuscript.

Funding: There is no external funding received for this research.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is publicly available at: https://www.kaggle.com/code/satyamsaini/deceptive-reviews/data (accessed on 15 May 2022).

Acknowledgments: This research was supported by the Researchers Supporting Project (RSPD2023R799), King Saud University, Riyadh, Saudi Arabia.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mahir, E.M.; Akhter, S.; Huq, M.R. Detecting fake news using machine learning and deep learning algorithms. In Proceedings of the 2019 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia, 28–30 June 2019; pp. 1–5.
2. Girgis, S.; Amer, E.; Gadallah, M. Deep learning algorithms for detecting fake news in online text. In Proceedings of the 2018 13th International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, 18–19 December 2018; pp. 93–97.
3. Toral, S.; Martínez-Torres, M.; Gonzalez-Rodriguez, M. Identification of the unique attributes of tourist destinations from online reviews. *J. Travel Res.* **2018**, *57*, 908–919. [CrossRef]
4. Jacobs, T.; Tschötschel, R. Topic models meet discourse analysis: A quantitative tool for a qualitative approach. *Int. J. Soc. Res. Methodol.* **2019**, *22*, 469–485. [CrossRef]
5. Popat, K.; Mukherjee, S.; Strötgen, J.; Weikum, G. CredEye: A credibility lens for analyzing and explaining misinformation. In Proceedings of the Web Conference 2018 (WWW '18), Lyon, France, 23–27 April 2018; pp. 155–158.
6. Agrawal, S.R. Adoption of WhatsApp for strengthening internal CRM through social network analysis. *J. Relatsh. Mark.* **2021**, *20*, 261–281. [CrossRef]
7. Racine, S.S.J. *Changing (Inter) Faces: A Genre Analysis of Catalogues from Sears, Roebuck to Amazon.com*; University of Minnesota: Minneapolis, MN, USA, 2002.
8. Skalicky, S. Was this analysis helpful? A genre analysis of the Amazon. com discourse community and its "most helpful" product reviews. *Discourse Context Media* **2013**, *2*, 84–93. [CrossRef]
9. Chen, C.; Wen, S.; Zhang, J.; Xiang, Y.; Oliver, J.; Alelaiwi, A.; Hassan, M.M. Investigating the deceptive information in Twitter spam. *Future Gener. Comput. Syst.* **2017**, *72*, 319–326. [CrossRef]
10. Feng, V.W.; Hirst, G. Detecting deceptive opinions with profile compatibility. In Proceedings of the sixth International Joint Conference on Natural Language Processing, Nagoya, Japan, 14–19 October 2013; pp. 338–346.
11. Cody, M.J.; Marston, P.J.; Foster, M. Deception: Paralinguistic and verbal leakage *Ann. Inter. Commu. Assoc.* **1984**, *8*, 464–490.
12. Ramalingam, A.; Navaneethakrishnan, S.C. An Analysis on Semantic Interpretation of Tamil Literary Texts. *J. Mob. Multimed.* **2022**, *18*, 661–682. [CrossRef]
13. Arenas-Marquez, F.J.; Martínez-Torres, M.R.; Toral, S. Electronic word-of-mouth communities from the perspective of social network analysis. *Technol. Anal. Strateg. Manag.* **2014**, *26*, 927–942. [CrossRef]
14. Govers, R.; Go, F.M. Deconstructing destination image in the information age. *Inf. Technol. Tour.* **2003**, *6*, 13–29. [CrossRef]
15. Conroy, N.K.; Rubin, V.L.; Chen, Y. Automatic deception detection: Methods for finding fake news. *Proc. Assoc. Inf. Sci. Technol.* **2015**, *52*, 1–4. [CrossRef]

16. Mondo, T.S.; Perinotto, A.R.; Souza-Neto, V. A user-generated content analysis on the quality of restaurants using the TOURQUAL model. *J. Glob. Bus. Insights* **2022**, *7*, 1–15. [CrossRef]
17. Perinotto, A.R.C.; Araújo, S.M.; Borges, V.d.P.C.; Soares, J.R.R.; Cardoso, L.; Lima Santos, L. The Development of the Hospitality Sector Facing the Digital Challenge. *Behav. Sci.* **2022**, *12*, 192. [CrossRef] [PubMed]
18. Santos, A.I.G.P.; Perinotto, A.R.C.; Soares, J.R.R.; Mondo, T.S.; Cembranel, P. Expressing the Experience: An Analysis of Airbnb Customer Sentiments. *Tour. Hosp.* **2022**, *3*, 685–705. [CrossRef]
19. Larcker, D.F.; Zakolyukina, A.A. Detecting deceptive discussions in conference calls. *J. Account. Res.* **2012**, *50*, 495–540. [CrossRef]
20. Barbado, R.; Araque, O.; Iglesias, C.A. A framework for fake review detection in online consumer electronics retailers. *Inf. Process. Manag.* **2019**, *56*, 1234–1244. [CrossRef]
21. Du, X.; Zhao, F.; Zhu, Z.; Han, P. DRDF: A Deceptive Review Detection Framework of Combining Word-Level, Chunk-Level, And Sentence-Level Topic-Sentiment Models. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 8–22 July 2021; pp. 1–7.
22. Weng, C.H.; Lin, K.C.; Ying, J.C. Detection of Chinese Deceptive Reviews Based on Pre-Trained Language Model. *Appl. Sci.* **2022**, *12*, 3338. [CrossRef]
23. Shojaee, S.; Murad, M.A.A.; Azman, A.B.; Sharef, N.M.; Nadali, S. Detecting deceptive reviews using lexical and syntactic features. In Proceedings of the 2013 13th International Conference on Intellient Systems Design and Applications, Salangor, Malaysia, 8–10 December 2013; pp. 53–58.
24. Olmedilla, M.; Martínez-Torres, M.R.; Toral, S. Harvesting Big Data in social science: A methodological approach for collecting online user-generated content. *Comput. Stand. Interfaces* **2016**, *46*, 79–87. [CrossRef]
25. Ku, Y.C.; Wei, C.P.; Hsiao, H.W. To whom should I listen? Finding reputable reviewers in opinion-sharing communities. *Decis. Support Syst.* **2012**, *53*, 534–542. [CrossRef]
26. Li, S.; Zhong, G.; Jin, Y.; Wu, X.; Zhu, P.; Wang, Z. A Deceptive Reviews Detection Method Based on Multidimensional Feature Construction and Ensemble Feature Selection. *IEEE Trans. Comput. Soc. Syst.* **2022**. [CrossRef]
27. Cao, N.; Ji, S.; Chiu, D.K.; Gong, M. A deceptive reviews detection model: Separated training of multi-feature learning and classification. *Expert Syst. Appl.* **2022**, *187*, 115977. [CrossRef]
28. Jacob, M.S.; Selvi Rajendran, P. Deceptive Product Review Identification Framework Using Opinion Mining and Machine Learning. In *Mobile Radio Communications and 5G Networks*; Springer: Singapore, 2022; pp. 57–72.
29. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2009.
30. hub, T. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available online: http://download.tensorflow.org/paper/whitepaper2015.pdf (accessed on 10 December 2022).
31. Ott, M.; Cardie, C.; Hancock, J.T. Negative deceptive opinion spam. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, 9–14 June 2013; pp. 497–501.
32. Ott, M.; Choi, Y.; Cardie, C.; Hancock, J.T. Finding deceptive opinion spam by any stretch of the imagination. *arXiv* **2011**, arXiv:1107.4557.
33. Rout, J.K.; Dalmia, A.; Choo, K.K.R.; Bakshi, S.; Jena, S.K. Revisiting semi-supervised learning for online deceptive review detection. *IEEE Access* **2017**, *5*, 1319–1327. [CrossRef]
34. Hassan, R.; Islam, M.R. Detection of fake online reviews using semi-supervised and supervised learning. In Proceedings of the 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox's Bazar, Bangladesh, 7–9 February 2019; pp. 1–5.
35. Etaiwi, W.; Awajan, A. The effects of features selection methods on spam review detection performance. In Proceedings of the 2017 International Conference on New Trends in Computing Sciences (ICTCS), Amman, Jordan, 11–13 October 2017; pp. 116–120.
36. Fusilier, D.H.; Montes-y Gómez, M.; Rosso, P.; Cabrera, R.G. Detection of opinion spam with character n-grams. In Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics, Cairo, Egypt, 14–20 April 2015; pp. 285–294.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.