Mining Behavioural and Sentiment-Dependent Linguistic Patterns from Restaurant Reviews for Fake Review Detection

Petr Hajek, University of Pardubice (CZ)

Jean-Michel Sahut, IDRAC Business School (FR)

Abstract. Online reviews are increasingly recognized as a key source of information influencing consumer behaviour. This in turn implies that competitive advantage can be achieved by manipulating users' perceptions about restaurants. The hospitality industry is particularly susceptible to this issue because products and services in this industry can only be rated upon consumption. Therefore, many efforts have recently been devoted to developing automatic methods for detecting fake reviews based on data intelligence in this sector. Recent studies suggest that both the semantic meaning of consumer reviews and the sentiment conveyed may be useful indicators of fake reviews. However, the semantic meaning may be context-sensitive and may also disregard sentiment information. Moreover, the content analysis approach should be integrated with the reviewer's behaviour to reveal their true intentions. To address these problems, we propose a review representation model based on behavioural and sentiment-dependent linguistic features that effectively exploit the domain context. Using a large dataset of Yelp restaurant reviews, we demonstrate that the proposed review representation model is more effective than existing approaches in terms of detection accuracy. It furthermore accurately estimates the average rating assigned by legitimate reviewers, which has significant managerial implications for the hospitality industry.

Keywords: fake review, detection, data intelligence, sentiment, behaviour, machine learning model, restaurant, hospitality industry.

1. Introduction

With the rapid development of consumer review websites, online travel reviews have become a central consideration for the hospitality industry by providing valuable information about product or service quality. Indeed, more and more tourists tend to share their experiences of restaurants, hotels and attractions via online platforms such as Yelp or TripAdvisor (Xiang et al., 2017; Hou et al., 2019). In a similar way to other high-involvement products, consumers are increasingly influenced by online travel reviews in their purchase decisions. This is demonstrated by consumers' willingness to search for other consumers' reviews and comments before making a purchase because these opinions are considered more trustworthy than advertisements disseminated by the businesses themselves (Schuckert et al., 2016). Perceived informativeness and persuasiveness of online reviews, together with source credibility and review volume, are regarded as the most important determinants of consumer purchasing intentions (Zhang et al., 2014). Notably, online reviews significantly affect revenues in the restaurant industry, with a one-star rating increase leading to a 5% to 9% increase in revenue (Luca, 2016). For firms in the hospitality industry, it is therefore becoming difficult to resist the temptation to produce or purchase (e.g., through freelance writers) fake (non-authentic) online reviews with either positive or negative polarity. This manipulation intends to promote the purchase of their products by publishing positive fake reviews or to use negative fake reviews to harm their competitors. The last decade has witnessed a remarkable increase in fake review volume in the hospitality industry, as evidenced by recent reports estimating that every third review on hotels or restaurants on TripAdvisor is fake (The Times, 2018)[1]. Moreover, fake online reviews tend to be more influential than genuine online reviews, urging the need to explore the features of fake reviews and develop methods to detect fake reviews (Wu et al.,

[1] https://www.thetimes.co.uk/article/hotel-and-caf-cheats-are-caught-trying-to-buy-tripadvisor-stars-027fbcwc8. Accessed 22 Jan 2020

2020; Paul and Nikolaev, 2021). Recent developments in text mining technology have led to an increased interest in automatic fake review detection (Cardoso et al., 2018; Vidanagama et al., 2020; Wu et al., 2020). Such automatic detection can be used to effectively monitor online platforms by providing higher accuracy than manual detection techniques (Ott et al., 2013). For example, Yelp uses its own automatic filter to issue a ranking penalty or even to remove a fraudulent user from their system altogether.

The last decade of research into fake review detection has seen the development of methods to automatically classify fake and genuine consumer reviews. This is a challenging problem because fake reviews are required to sound authentic as if written by real consumers. Three main types of methods were proposed to tackle this problem, namely, machine learning, network-based and pattern-mining approaches (Vidanagama et al., 2019). The most significant results have been achieved using machine learning approaches based on review- and reviewer-centric features (Wu et al., 2020). The content of the review is analysed in the review-centric approaches. However, methods based purely on the content of reviews (linguistic features) can be easily manipulated by spammers, e.g. by rewording the content of genuine reviews. It is therefore desirable to combine the review-centric features with those related to reviewer behaviour and characteristics because spammers tend to write more frequent, shorter and more positive reviews than legitimate reviewers (Crawford et al., 2015). The main limitation of machine learning approaches, however, is that large manually-labelled data are required to achieve competitive accuracy. Machine learning methods also struggle to deal with high-dimensional and sparse data obtained using the traditional bag-of-words approach (selection of the most frequently appearing words and phrases). Moreover, the semantic meaning of consumer reviews cannot be directly identified using standard word counts. To address these problems, neural networks (NNs) have been employed in recent years to produce a lower dimensional and dense word representation (Mitra and Jenamami, 2021). Dimensions in such representations correspond to particular word contexts, thus capturing useful semantic and syntactic features (Ren and Ji, 2017; Li et al., 2017; Hajek et al., 2020). However, this approach also presents several drawbacks. First of all, a large set of reviews must be considered to effectively exploit the word contexts. Secondly, the word representations generated are context-sensitive, which in turn requires a large number of documents from a specific product domain. Thirdly, as indicated above, consumer sentiment may be important for fake review detection. However, word embeddings disregard sentiment information. Recently, Martinez-Torres and Toral (2019) demonstrated that sentiment-related unique attributes result in non-biased detection of fake reviews in the hospitality industry. More precisely, several bag-of-words were selected separately for positive and negative fake and legitimate reviews to improve detection performance. Inspired by this finding, here we adopt a unique approach of combining behavioural and sentiment-dependent linguistic patterns to detect fake reviews to enable the accurate estimate of the average rating assigned by legitimate reviews. More precisely, we propose a supervised feature selection algorithm to extract sentiment-based bag-of-words and, additionally, adapt the pre-trained word embeddings to consider the sentiment of reviews. Hence, in summary, the contributions are threefold:

- We introduce two novel sets of sentiment-dependent textual features to identify linguistic patterns in fake review data. This is, to the best of our knowledge, the first time bag-of-words and word embedding representations are modified in order to capture sentiment information.
- To overcome the problem of naturally skewed distribution of the real-world fake review datasets towards the legitimate reviews, we apply a clustering-based under-sampling approach here, which utilizes the underlying data more effectively compared with random approaches used previously.

- The fusion of the linguistic patterns and behavioural patterns extracted from online reviews enables a novel fake review detection model to be developed. Using a benchmark dataset of Yelp restaurant reviews, we demonstrate the effectiveness of the model compared with state-of-the-art models by considering the extent to which the model corrects the detrimental effect of fake reviews on average review ratings.

The remainder of the paper is structured as follows. Section 2 presents the theoretical framework and hypothesis. Section 3 exposes the research methodology, including the description of the data used for empirical validation. Section 4 analyses the results of data pre-processing and fake review detection. Section 5 discusses the results obtained and presents implications. Finally, section 6 concludes.

2. Theoretical Framework and Hypotheses

2.1 Fake Review Detection in the Hospitality Industry

Fake reviews have increasingly been recognized as a major concern in hospitality management. Fake reviews are intended to achieve competitive advantage by promoting or demoting target services and products (Ren and Ji, 2017). In other words, fake reviews are produced to influence travellers' consumer decisions. Indeed, travellers trust online reviews and have a limited capacity to identify fake reviews (Heydari et al., 2015). Automated methods based on data intelligence have been used to ensure their early detection by classifying reviews into fake and legitimate categories. Regarding the features used in previous studies and summarized in Table 1, review-centric and reviewer-centric features were exploited by mainly considering textual and behavioural features of reviews and reviewers, respectively. Among the most representative studies, Ott et al. (2013) used standard bag-of-words (*n*-grams) by calculating term and document frequencies of words or phrases. However, this document representation presents the weakness of high dimensionality and sparsity, which makes accurate detection of fake reviews difficult for traditional machine learning methods. Moreover, such representation disregards the context and semantics of words. Therefore, textual features were extended by considering psycholinguistic features obtained using word lists (dictionaries), such as positive / negative word lists (Mukherjee et al., 2013), subject words (Deng and Chen, 2014), spatial words (Li et al., 2014), and cognitive, social and perceptual words (Li et al., 2020). In addition, part-of-speech (POS) tagging was performed to consider the fact that fake reviews include more first-person pronouns (Li et al., 2014). It was also found that fake reviews tend to include more verbs than nouns and that fake reviews lack temporal references and structure (Plotkina et al., 2020). Other lexical and syntactic features used in previous studies include occurrences of punctuation and vocabulary richness (Shojaee et al., 2013) and occurrences of capital letters (Rayana and Akoglu, 2015). To overcome the above-mentioned limitations of bag-of-words features, word embeddings were introduced to consider word context and to produce low-dimensional dense review representations. Ren and Ji (2017) used the pre-trained word embeddings to encode the semantic meaning of review sentences. Google's BERT (Bidirectional Encoder Representations from Transformers) was used to fine-tune pre-trained word representations, resulting in a model with sentential context (Kennedy et al., 2019). To capture the weights of *n*-grams in the semantic representation, Li et al. (2019) utilized a recurrent deep neural network with attention mechanism. However, these approaches share a common disadvantage in that they fail to consider sentiment information. To overcome this deficiency, some recent studies combine the above-mentioned sentiment features based on positive and negative word lists with the word embedding features (Kennedy et al., 2019; Liu et al., 2019; Hajek et al., 2020).

In addition, detecting fake reviews using only textual features is a challenging task because fake reviews are generally well written (Plotkina et al., 2020). Therefore, textual features should be accompanied with behavioural features of the reviewer to reveal their true intentions.

Such features refer to the reviewer’s activity, including their rating features, popularity and spatial information (Barbado et al., 2019; Li et al., 2020; Ruan et al., 2020). Compared with legitimate reviews, fake reviews tend to have more polarized distribution (Luca and Zervas, 2016). Product features and ratings were also incorporated in previous studies, such as their temporal aspects and burstiness of reviews (Rayana and Akoglu, 2015). The presence of concept drift was also investigated in the stream of reviews (Mohawesh et al., 2021). The difference between the specific review and average rating for the product was also used as an indicator of fake reviews (Schuckert et al., 2016; Shan et al., 2021).

Methods used for fake review detection include graph-based and machine learning methods. Graph-based methods are used to capture relationships among products, stores, reviews and reviewers in a graph by representing each node of the graph with a set of features (Shehnepoor et al., 2017; Fang et al., 2020; Manaskasemsak et al., 2021). Machine learning methods can be further categorized into supervised, semi-supervised and unsupervised learning. Machine learning methods with supervised learning, such as support vector machines (SVM) and neural networks, are reportedly most accurate but require reliable labels of fake and legitimate reviews. Manual labels are not considered reliable because people are generally not accurate in detecting fake reviews (Plotkina et al., 2020). Therefore, anonymous online workers were asked to pretend to act as customers and produce realistic fake reviews (Ott et al., 2013). In this manner, gold-standard datasets were created to enrich detection methods in the hospitality domain (Li et al., 2014). The main disadvantage of this data generation is that randomly chosen online workers do not possess sufficient experience and domain knowledge to produce convincing fake reviews (Vidanagama et al., 2020). Moreover, only a limited number of fake reviews can be generated in this way, which in turn has a detrimental effect on the detection performance of machine learning methods. To address this issue, semi-supervised and unsupervised learning methods take advantage of incorporating unlabelled data. Using unlabelled data together with a small sample of labelled data was particularly effective in previous studies (Rayana and Akoglu, 2015; Yilmaz and Durahim, 2018). Another solution is to collect larger real-life datasets filtered by commercial websites, such as Yelp or Amazon. Indeed, existing literature suggests that such review filtering is reasonable and accurately detects fake review activity (Mukherjee et al., 2013; Kennedy et al., 2019; Li et al., 2020).

Table 1: Summary of features and methods used in previous studies

Study	Review-centric / Reviewer-centric features	Method	Data (source)	# fake / legitimate	Performance
Ott et al. (2013)	*n*-grams	SVM	Hotels	800/800	F-score=0.884
Shojaee et al. (2013)	occurrences of punctuation, vocabulary richness, character counts	SVM	Hotels	800/800	F-score=0.840
Mukherjee et al. (2013)	*n*-rams, POS, LIWC	SVM	Yelp	Hotel 802/4,876 Restaurant 8,368/50,149	F-score=0.692 F-score=0.711
Deng and Chen (2014)	positive and negative words, subject words	NB	Restaurant (Dianping)	17,681/38,802	Acc=0.740
Li et al. (2014)	unigrams, positive and negative words, spatial words, first-person pronouns	SAGE	Hotels, restaurants (Chicago hotels and restaurants)	800/800, 200/200	F-score=0.784
Li et al. (2015)	*n*-grams / registered user, distance from Shanghai, rating deviation, # of unique IPs, cookies and cities	SVM	Restaurants (Dianping)	~ 6.1 mil. reviews	F-score=0.850
Rayana and Akoglu (2015)	capital letters, review length, first-person pronouns, subjective and objective words, description length / review rank order, rating deviation, review time	SSL	YelpChi (Chicago) YelpNYC (New York) YelpZip (New York, Vermont, Connecticut, Pennsilvania)	8,916/58,479 36,875/322,177 80,457/528,141	AUC=0.789 AUC=0.770 AUC=0.794
Sun et al. (2016)	product word embeddings, *n*-grams	Bagging, SVM, CNN	Amazon	800/1,200	F-score=0.772
Luca and Zervas (2016)	review length, rating / # of reviews, user photo, # of friends	-	Yelp (Boston restaurants)	50,486/265,929	R^2=0.430
Li et al. (2017)	sentence weights, POS, first-person pronouns	CNN, SWNN	Hotels, restaurants	800/800, 200/200	F-score=0.861
Ren and Ji (2017)	word embeddings	CNN, GRNN	Hotels, restaurants	800/800 200/200	F-score=0.774 F-score=0.870
Yilmaz and Durahim (2018)	word embeddings / reviewer-product network	SSL	Yelp	8,916/58,479 36,875/322,177 80,457/528,141	AUC=0.807 AUC=0.813 AUC=0.832
Ahmed et al. (2018)	*n*-grams	SVM	Hotels	800/800	Acc=0.870
Zeng et al. (2019)	first sentence, middle context, last sentence	LSTM ensemble	Hotels, restaurants	800/800 200/200	F-score=0.857 F-score=0.832
Barbado et al. (2019)	bigrams, positive and negative words / rating deviation, real name, bookmarks, registration date, votes received, content similarity, # of friends and followers, # of photos, # of reviews, rating distribution	AdaBoost	Yelp (New York, Los Angeles, Miami, San Francisco)	9,456/9,456	F-score=0.810
Kennedy et al. (2019)	word embeddings, review length, capitalized words, numerals, POS, positive and negative words	BERT	Hotels, YelpZip	800/800 80,456/528,142	F-score=0.888 F-score=0.731
Liu et al. (2019)	positive and negative words, review length, first-person pronouns, multimodal embeddings	LR	Hotels, restaurants (Dianping)	16,044/15,273 47,246/50,593	F-score=0.790 F-score=0.820
Barushka and Hajek (2019)	word embeddings, *n*-grams	DFFNN	Hotels	400/400	Acc=0.891
Martinez-Torres and Toral (2019)	polarity-oriented unique *n*-grams	SVM, *k*-NN, LR, RF, GB, MLP	Hotels	800/800	F-score=0.881
Hajek et al. (2020)	word embeddings, *n*-grams, lexicon-based emotions	DFFNN, CNN	Hotels, restaurants	400/400 200/200	F-score=0.896 F-score=0.901
Li et al. (2020)	positive and negative words, review length, rating, time difference, cognitive, social and perceptual words / # of reviews, # of friends, reviewer location	LR	Yelp	6,754/36,742	R^2=0.128
Ruan et al. (2020)	geolocation features, account features	GADM	Yelp	20,267/87,357	F-score=0.862

Shan et al. (2021)	language style, behavioural features, rating-sentiment inconsistency, content inconsistency, language inconsistency	RF, DT, NB, SVM, MLP	Yelp	11,641/12,898	F-score=0.932

Legend: Acc – accuracy, AUC – area under ROC curve, BERT – bidirectional encoder representations from transformers, CBUS – clustering-based under-sampling, CNN – convolutional neural network, DFFNN – deep feed-forward neural network, DT – decision tree, F-score – F1-score (average of precision and recall), GADM – geolocation-based account detection model, GB – gradient boosting, GRNN – general regression neural network, *k*-NN – *k*-nearest neighbour, LIWC – linguistic inquiry and word count, LR – logistic regression, LSTM – long short term memory, MLP – multilayer perceptron, NB – Naïve Bayes, POS – part-of-speech tagging, RF – random forest, SAGE – sparse additive generative model, SSL – semi-supervised learning, SVM – support vector machine, SWNN – sentence weighted neural network.

2.2 Hypothesis Development

As noted above, results of earlier research indicate that reviews with different sentiment polarization (positive and negative) have different linguistic structures (Plotkina et al., 2020). Moon et al. (2021) report that fake reviews tend to contain extreme positive and negative expressions. Moreover, fake reviews are prone to more polarized rating distribution (Luca and Zervas, 2016). Existing research also indicates that positive fake reviews prevail over negative fake reviews (Zhang, 2019). Therefore, fake reviews with different polarity were investigated separately (Ott et al., 2013), which led to biased classification models (Martinez-Torres and Toral, 2019). Generally, detecting negative fake reviews is reportedly more difficult than positive reviews (Fusilier et al., 2015). Hence, sentiment analysis became a critical tool for mining textual features from fake reviews (Barbado et al., 2019; Liu et al., 2019, Kennedy et al., 2019; Li et al., 2020; Hajek et al., 2020). Traditional approaches to sentiment analysis of fake reviews were based on calculating positive and negative word counts. Positive and negative word lists must be available for such sentiment analysis. For instance, the LIWC (linguistic inquiry and word count) tool was frequently used to calculate the sentiment scores (Mukherjee et al., 2013; Li et al., 2014; Li et al., 2020). Similarly, the HowNet sentiment dictionaries were applied for Chinese fake review datasets (Liu et al., 2019). However, using positive and negative sentiment scores based on a single dictionary leads to limited lexical coverage and unreliable sentiment analysis (Bravo-Marquez et al., 2014). To address these issues, a combination of different sentiment dictionaries (lexicons) was proposed (Hajek et al., 2020). Broadly speaking, the lexicon-based approaches have several major limitations (Dhaoui et al., 2017). First of all, standard dictionaries are compiled manually and not specifically for a particular domain. Furthermore, all words included in the dictionaries are considered equally important or weightings are assigned to only a few of them. However, the same words or phrases may have different sentiment polarity and weighting across domains. Moreover, emoticons and emojis, abbreviations, colloquialisms and misspellings are not covered. To remedy these deficits, machine learning methods are used to learn the word list (bag-of-words) automatically without relying on a pre-defined dictionary. A set of documents categorized into positive and negative classes must be available for such approach. Consequently, words and phrases (*n*-grams) are extracted from the documents based on the frequencies of their occurrence. To calculate the frequencies, a term-weighting scheme is used by combining term frequency with the inversed document frequency of the term across the set of documents. Hence, high relevance is given to words and phrases that occur frequently in a limited number of documents. Finally, top ranked features (*n*-grams) are selected in terms of their relevance. However, this traditional procedure does not capture sentiment information. Martinez-Torres and Toral (2019) used ANOVA for filtering *n*-grams with significantly higher discriminative power when distinguishing positive and negative fake and legitimate reviews. On the one hand, this leads to dimensionality reduction, which is considered important for traditional machine learning methods. On the other hand, sentiment information is not reflected in the weightings of selected *n*-grams. In the sentiment analysis literature (Deng et al., 2014), extensions of the traditional term weighting schemes were introduced to improve the accuracy of sentiment analysis. The modified supervised weighting scheme considers both the importance of a term in a document and its importance for expressing sentiment. Here we use this method to achieve more efficient feature selection that incorporates sentiment information. So, the following hypothesis is formulated:

H1. *A sentiment-dependent supervised term weighting scheme is more effective for selecting bag-of-words than a traditional term weighting scheme and positive / negative sentiment scores based on word lists.*

In a similar way to the bag-of-words representation, the dense low-dimensional representation obtained using word embeddings disregards sentiment information in previous studies (Ren and Ji, 2017; Yilmaz and Durahim, 2018). To capture this information, here we fine-tune the pre-trained word embeddings using the sentiment polarity of reviews in training data. More precisely, we use a convolutional neural network (CNN) to perform sentiment analysis, and then extract the learned features from the input of the dense layer of the trained CNN model. It is expected that sentiment-dependent word embeddings will improve the performance of fake review detection. Therefore, the following hypothesis is formulated:

H2. *Word embeddings fine-tuned using a sentiment classifier outperforms the pre-trained word embeddings.*

Although textual features were successfully used in earlier research on fake review detection, using them separately from behavioural data limits the detection accuracy of machine learning methods (Rayana et al., 2015). Moreover, richer behavioural features are reportedly more accurate for real-world (commercial) fake review datasets (Mukherjee et al., 2013; Hussain et al., 2020; Wang et al., 2020). This is because the authors of fake reviews often share activity patterns and profile characteristics (Crawford et al., 2015): (1) they produce similar reviews for different products; (2) their ratings deviate from the average rating provided by legitimate reviewers; (3) they usually produce shorter reviews; (4) they generate many reviews within a short period of time; and (5) they produce a high percentage of positive (five-star) reviews. Therefore, since this research investigates a real-world restaurant fake review dataset, we can expect than behavioural patterns extracted from the data will be more effective than their textual counterparts. To further improve the performance of fake review detection, we propose to integrate the two sources of data and develop a neural network detection model that captures their high-level features. We also focus on the detrimental effects of fake reviews. Unlike previous studies, here we consider the consequences of fake reviews not only regarding their detection accuracy but also in terms of their effect on the average rating of the product. In summary, the following hypothesis is formulated:

H3. *The combination of behavioural and sentiment-dependent linguistic patterns increases the accuracy of fake review detection and improves the estimate of the average rating assigned by legitimate reviews.*

3. Methodology

In order to construct the machine learning model with supervised learning, novel sentiment-dependent textual features are extracted from the review dataset and combined with behavioural patterns to effectively integrate the review-centric and reviewer-centric features.

3.1 Sentiment-Dependent Linguistic Features

To obtain sentiment-dependent linguistic features, we extracted two sets of features, namely sentiment-dependent bag-of-words and sentiment-dependent word embeddings.

To perform sentiment-dependent feature extraction, reviews must be categorized into sentiment classes. For this purpose, let us first introduce some basic notations. Let D^{pos} and D^{neg} be the sets of reviews of positive and negative sentiment classes, respectively. Following previous studies (Moraes et al., 2013; Ott et al., 2013), reviewers' ratings were used for the categorization so that four- and five-star ratings indicated positive sentiment while one- and two-star ratings denoted negative sentiment.

In the bag-of-words representation, each review r_j is defined as r_j = ($w_{\mathrm{j}1}$, w_{j2}, … , w_{jm}), where the vector of weightings represents the importance of terms t_1, t_2, … , t_m. To reduce noise in the data, we removed stopwords using the Rainbow algorithm, transformed all words to lowercase letters and tokenized them to identify *n*-grams (unigrams, bigrams and trigrams). In the weighting scheme used, term frequency ITD(t_i, r_j) of term t_i in review r_j is combined with the capacity ITS(t_i) of t_i to discriminate positive and negative sentiment as follows (Deng et al., 2014):

$$w_{ij} = \mathrm{ITD}(t_i, r_j) \times \mathrm{ITS}(t_i), \tag{1}$$

where ITD(t_i, r_j) is obtained as the raw frequency normalized by the length of review, and ITS(t_i) is calculated as the maximum value of WFO (weighted frequency and odds) over the two sentiment categories. WFO can be estimated as follows:

$$WFO(f_i, D^k) \approx \left(\frac{x_i^k}{N_k}\right)^{\lambda} \log\left(\frac{x_i^k(N_1+N_2+N_k)}{y_i^k N_k}\right)^{1-\lambda}, \tag{2}$$

where x_i^k and y_i^k respectively denote the number of reviews belonging to D^k (D^{pos} or D^{neg}) and not belonging to D^k that contain term t_i; N^k is the number of reviews in D^k; and parameter λ is the ratio between frequency and odds. Finally, terms were ranked according to their w_{ij} and, in agreement with earlier studies (Kouloumpis et al., 2011), the top 1,000 terms were selected as the bag-of-words features. Unlike the traditional *tf.idf* (document frequency – inverse document frequency) weighting scheme, the set of terms selected using the supervised term weighting scheme accounts for differences between the positive and negative sentiment classes of reviews.

To generate sentiment-dependent word embeddings, we used the pre-trained 300-dimensional Glove word embeddings (Pennington et al., 2014) originally trained on a dataset of 42 billion words with a vocabulary of 1.9 million words[2]. Then, the CNN model was created to fine-tune the embedding weight matrix on the review dataset categorized into positive and negative sentiment class. For the input layer, reviews were treated as word sequences {w_1, w_2, … , w_T} with words w_t drawn from a vocabulary *V*. Word vectors $\boldsymbol{w} \in \mathbb{R}^{1\times d}$ are looked up in a word

[2] https://nlp.stanford.edu/projects/glove/

embedding matrix $\boldsymbol{W} \in \mathbb{R}^{|V| \times d}$. A review matrix $\boldsymbol{R} \in \mathbb{R}^{T \times d}$ is generated for each review r, where T is the maximum number of words in the reviews (maximum length).

One convolutional layer followed with 50 filters of size 5 (using rectified linear unit (ReLU) activation functions). A max-pooling layer was added with a pool size of 4. Flattening is performed to convert the feature map from 2D max-pooling into a one-dimensional array used as an input to a dense (fully-connected) layer with 100 sigmoid units. The output layer calculates the probability distribution over the positive and negative sentiment classes. The proposed CNN architecture is depicted in Fig. 1.

Adam optimizer was used for learning the CNN model with 5 epochs and cross-entropy loss as the fitness function. The experiments were performed using Keras library on a Jetson AGX Xavier Developer Kit equipped with 512-core Volta GPU with Tensor Cores and 32GB memory.

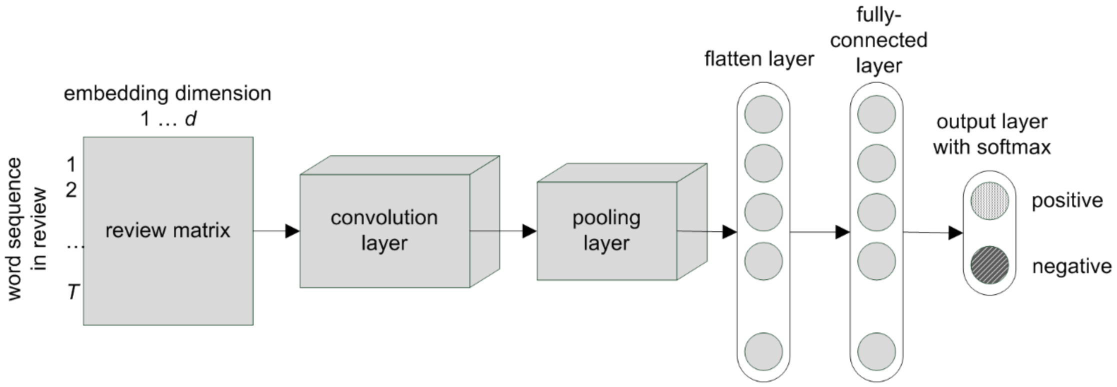


Fig. 1 The CNN architecture for extracting sentiment-dependent word embeddings

3.2 Behavioural Features

According to the literature, abnormal reviewer behavioural patterns likely to be related to fake reviews are based on the following list of features:

(1) *Average rating* given by the reviewer avg(☆r_a), where ☆r_a denotes a rating on review r by author a on the 5-star rating scale, to indicate whether the reviewer gives an abnormal percentage of positive (negative) reviews (Jindal and Liu, 2007). *Extreme*

rating (1☆ or 5☆) was used to represent a spammer's intention to demote or promote product *p* (Mukherjee et al., 2013). The reviewer's *average rating deviation* from their average rating was also included to identify abnormal rating behaviour (Jindal and Liu, 2007).

(2) *Rating deviation* to consider spammers' intentions to manipulate average rating for product *p*. This feature is calculated as the average for an author over their absolute rating deviations of a review r_a from other reviews on the same product $p(r_a)$. Previous studies found that legitimate reviewers show substantially lower rating deviation compared with spammers (Mukherjee et al., 2013).

(3) *Early time frame* measured as the number of days since the first review to detect early reviews that greatly affect customers' perceptions (Mukherjee et al., 2013b; Hussain et al., 2020). The *rank order* (sorted by date) among all the reviews *r* for product *p* is also calculated (Jindal and Liu, 2007).

(4) Abnormal numbers of reviews posted by a single reviewer and for a single product are also considered to detect review burstiness. Specifically, the *maximum number of reviews* per day for author *a*, *maximum number of reviews* per day for product *p*, and the overall number of reviews posted by author *a* are calculated. The latter feature is used to detect authors that are not long-time members because spammers tend to post a lower number of reviews compared to legitimate reviewers (Hussain et al., 2020).

(5) *Review length* to consider the spammers' lack of product experience and reluctance to spend much time on writing reviews. Empirical results confirmed that average review length for spammers is generally shorter than that for legitimate reviewers (Mukherjee et al., 2013b).

3.3 Integrated Fake Review Detection Model

The natural class distribution of the real-life review datasets is skewed in favour of legitimate reviews. This data imbalance in turn leads to poor detection performance of machine learning models due to bias towards the legitimate class (Mukherjee et al., 2013; Crawford et al., 2015). To overcome this problem, data sampling methods were used which produce balanced data. Mukherjee et al. (2013) and Liu et al. (2019) used random under-sampling to match the number of fake reviews by selecting a random subset of legitimate reviews. On the one hand, under-sampling seems to be an appropriate approach to deal with imbalanced fake review data due to sufficient real-world fake reviews and, unlike over-sampling, under-sampling does not suffer from overfitting issues (Budhi et al., 2021). On the other hand, potentially important reviews can be removed from the legitimate class using random over-sampling. To avoid this major drawback, the clustering-based under-sampling method was used (Yen and Lee, 2006), which selects the subset of the legitimate class reviews from each cluster based on the ratio of the numbers of legitimate and fake reviews in the cluster. Following Yen and Lee (2006), the number of clusters was set to three and the algorithm performed in the Keel Suite 3.0 environment. This method is robust to the number of clusters, but a relatively small number of clusters (two to four) is recommended to avoid overfitting (Yen and Lee, 2009; Ofek et al., 2017). Hence, fake and legitimate classes were balanced in the training data.

To predict fake / legitimate classes in the review dataset, this paper proposes an integrated neural network model which learns high-level features from the behavioural and sentiment-dependent linguistic patterns. A neural network architecture comprising two dense (fully-connected) layers with ReLU units (Fig. 2) was trained using the mini-batch gradient descent algorithm (a stable convergence was achieved using the following setting: the number of mini-batches set to 100, learning rate to 0.1, and the number of iterations to 500). To avoid overfitting, dropout regularization was applied for each layer (dropout rate = 0.2 for the input layer, dropout rate = 0.5 for the hidden layers). For the objective function, cross-entropy loss was used, and the experiments were also performed using Keras on the Jetson AGX Xavier Developer Kit. It should be noted that different hyper-parameters were experimented (e.g., numbers of dense layers and units in the hidden layers) but due to space constraints only the results for the best-performing architecture are presented in the next section.

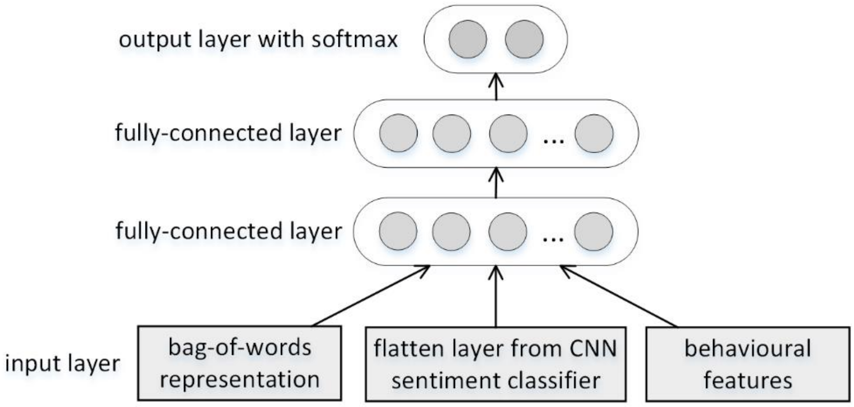

output layer with softmax
fully-connected layer
...
fully-connected layer
...
input layer
bag-of-words representation
flatten layer from CNN sentiment classifier
behavioural features

Fig. 2 The proposed neural network architecture for fake review detection

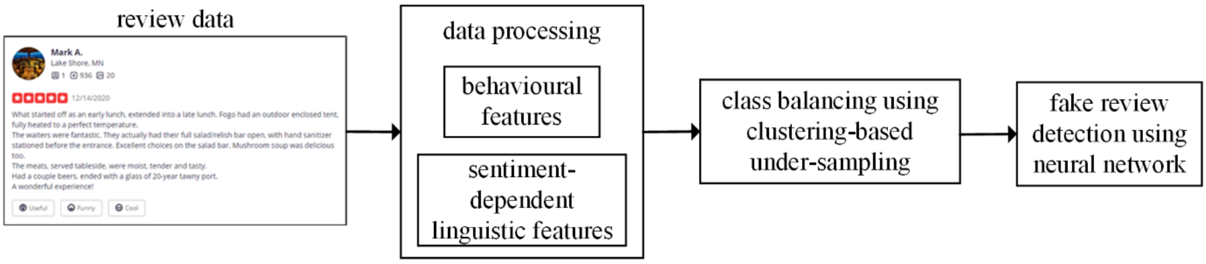
To evaluate the performance of the proposed model, two measures were used. Area under the receiver operating characteristic curve (AUC) represents a standard evaluation measure for skewed data due to its robustness against class imbalance. The AUC corresponds to the probability that the machine learning method ranks a randomly chosen fake review higher than a random legitimate review. From the perspective of a two-class classification problem, the AUC represents a trade-off between the true-positive and false-positive rate. The second evaluation measure is the deviation of the detection model (classifier) from the average rating assigned by legitimate reviewers. The rating deviation of the classifier can be defined as follows:

$$r^{dev} = \frac{1}{K}\sum_{k=1}^{K}|\hat{p}_k(r) - p_k(r)|, \qquad (3)$$

where $p_k(r)$ is the average rating assigned to the k-th product (store) by legitimate reviewers, $\hat{p}_k(r)$ represents the average rating obtained after removing reviews classified as fake, and K

is the number of products. This measure enables us to evaluate to what extent the fake review detection model reduces the detrimental effect of fake reviews on average product rating.

All components of the proposed fake review detection model are depicted using a simplified conceptual framework in Fig. 3.


review data
Mark A.
Lake Shore, MN
1 936 20
12/14/2020
Useful
Funny
Cool
data processing
behavioural features
sentiment-dependent linguistic features
class balancing using clustering-based under-sampling
fake review detection using neural network

What started off as an early lunch, extended into a late lunch. Fogo had an outdoor enclosed tent, fully heated to a perfect temperature.
The waiters were fantastic. They actually had their full salad/relish bar open, with hand sanitizer stationed before the entrance. Excellent choices on the salad bar. Mushroom soup was delicious too.
The meats, served tableside, were moist, tender and tasty.
Had a couple beers, ended with a glass of 20-year tawny port.
A wonderful experience!

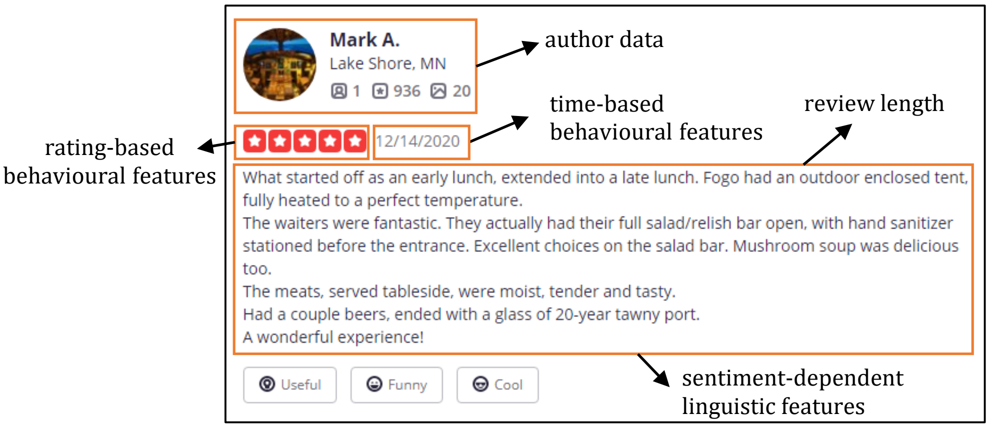
Fig. 3 Conceptual framework of the proposed fake review detection model

3.4 Data

As the recent literature shows that the filtering of reviews by commercial websites accurately detects fake review activity (Kennedy et al., 2019), the YelpZip benchmark dataset was used, collected from Yelp.com and first presented by Rayana and Akoglu (2015)[3]. The authors of this dataset searched for restaurants by zipcode to include those located in NY and neighbouring states CT, NJ, VT, and PA. The dataset comprises 608,598 reviews for 5,044 restaurants posted by 260,277 reviewers. In addition to review text, reviews also included user and product information, timestamp, and rating on a five-star scale (see Fig. 4). To label the reviews by fake / legitimate class, the filtered list of fake reviews is used as provided by the filtering algorithm of Yelp. As indicated above, this algorithm is considered not perfect but accurate enough to detect fake review activity (Mukherjee et al., 2013). To label legitimate reviews, the list of recommended reviews was used. To extract behavioural patterns, authors of fake reviews were categorized as spammers (23.91%) while authors with no fake reviews

[3] http://odds.cs.stonybrook.edu/yelpzip-dataset/

were labelled non-spammers. Of 608,598 reviews in the dataset, 80,457 (13.22%) were fake reviews and 528,141 were legitimate reviews.



Mark A.
Lake Shore, MN
1 936 20
12/14/2020

What started off as an early lunch, extended into a late lunch. Fogo had an outdoor enclosed tent, fully heated to a perfect temperature.
The waiters were fantastic. They actually had their full salad/relish bar open, with hand sanitizer stationed before the entrance. Excellent choices on the salad bar. Mushroom soup was delicious too.
The meats, served tableside, were moist, tender and tasty.
Had a couple beers, ended with a glass of 20-year tawny port.
A wonderful experience!

author data
time-based behavioural features
review length
rating-based behavioural features
Useful
Funny
Cool
sentiment-dependent linguistic features

Fig. 4 Illustration of textual and behavioural features in a Yelp review

4. Experimental Results

In this section, the results of the experiments are presented to evaluate the effectiveness of the proposed sentiment-dependent linguistic features for fake review detection on the YelpZip restaurant dataset. To evaluate the performance of fake review detection, a 80/20 stratified split for training and testing data was used. Hence, the class prevalence was maintained between data splits. Reliable evaluation was achieved by repeating this split procedure ten times. Hereinafter, the mean values of the evaluation measures are reported for the testing set.

To extract the sentiment-dependent linguistic features, unigrams, bigrams and trigrams were first identified in the pre-processed textual data. Then, their weightings w_{ij} were calculated using Eq. (1) and the top 1,000 ranked *n*-grams were selected to produce the sentiment-dependent bag-of-words features. Recall that these features consider not only term frequencies but also their discriminative power for sentiment classification. Table 2 presents thirty top ranked features according to the proposed weighting scheme.

Table 2: Top ranked sentiment-dependent bag of words features

a great	but the	Enjoyed	Friendly	love	perfect
amazing	delicious	Fantastic	food was	minutes	restaurant
as well	didn’t	Favorite	Give	ordered	spot
awesome	don’t	food was	Great	out of	the best
bad	excellent	Fresh	Happy	people	wonderful

To demonstrate the effect of the sentiment-dependent bag of words features on fake review detection, the neural network architecture introduced above was used for the integrated model (i.e., with two dense fully-connected layers of 100 and 50 ReLU units trained using the mini-batch gradient descent algorithm with dropout regularization). These features were compared with the baseline obtained for the bag-of-words features selected as the top 1,000 *n*-grams according to the traditional *tf.idf* weighting scheme. Fig. 5 shows that the mean values of AUC and weighted F-score were improved when using the sentiment-dependent bag-of-words (BoW) features, indicating the effectiveness of the enhanced weighting scheme in fake review detection.

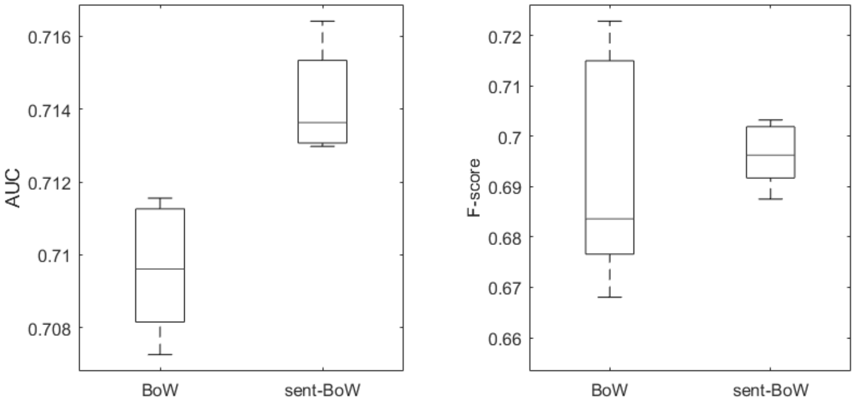


Fig. 5: The effect of the BoW features on the detection model performance

Note: Neural network with two hidden layers of 100 and 50 neurons was trained on the sentiment-dependent BoW representation (sent-BoW) compared with the traditional BoW representation based on *tf.idf* weights.

To produce the sentiment-dependent word embeddings (SWE), the pre-trained 300-dimensional Glove word embeddings were fine-tuned using the CNN model presented in Fig. 1. The mean value of AUC for the CNN-based sentiment classifier was 0.8829. Then, the output of the flattened layer from the trained CNN model was used as the input layer of the neural network with two dense layers, the same way as for the bag-of-words representation. Again, a baseline of the pre-trained word embeddings was applied to demonstrate the improvement in detection performance. Fig. 6 shows that fake review detection can be improved for the fine-tuned word embedding features.

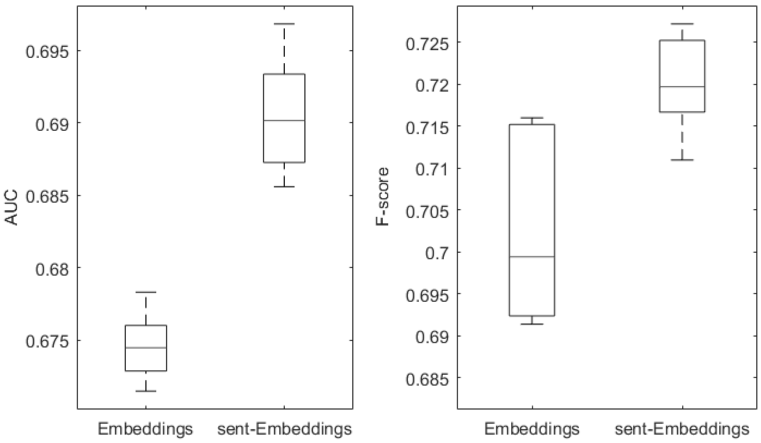


Fig. 6: The effect of the SWE features on the detection model performance

Note: A neural network with two hidden layers of 100 and 50 neurons was trained using the flattened layer of the CNN-based sentiment classifier (sent-Embeddings) compared with the pre-trained word embeddings averaged per review (Embeddings).

The above results indicate the effectiveness of the sentiment-dependent linguistic features. In a further set of experiments, the synergic effects of combining these features with behavioral features into an integrated fake review detection model were investigated.

To investigate the effect of balancing the dataset using the clustering-based under-sampling algorithm, the performance of the proposed fake review detection model was compared against that obtained for the original imbalanced dataset, as depicted in Fig. 7. The results show a noticeable increase in the performance measures, indicating that the under-sampling procedure effectively selected relevant data instances. In addition, we demonstrate that the clustering-based under-sampling algorithm outperforms the SMOTE (synthetic minority oversampling technique) algorithm, indicating overfitting of the over-sampling algorithm, which is consistent with other recent findings on this dataset (Budhi et al., 2021). Furthermore, we performed overfitting analysis to check whether the neural network architecture is adequate. Fig. 8 shows that the best results were obtained by using a neural network with two hidden layers and that in terms of AUC, the model performed best with 100 and 50 neurons, respectively.

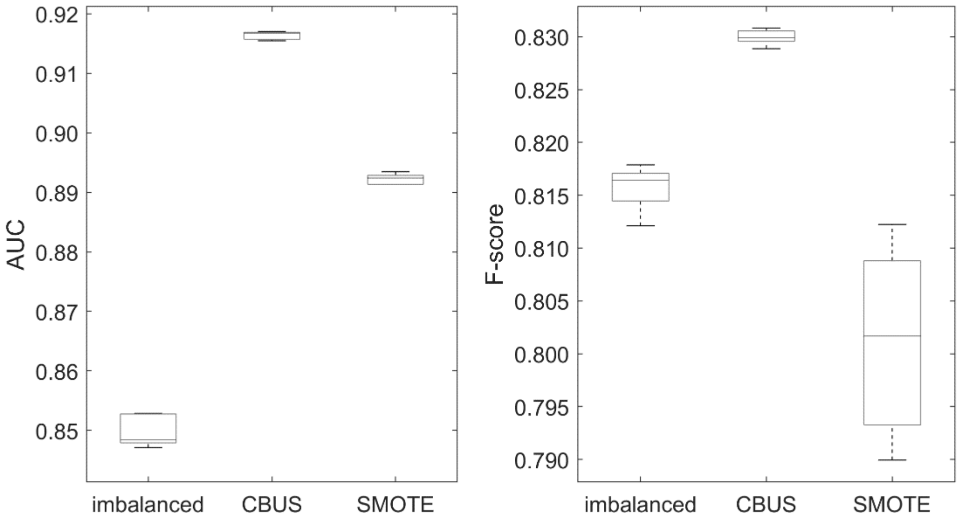


Fig. 7: The effect of sampling methods on the detection model performance

Note: A neural network with two hidden layers of 100 and 50 neurons was trained for a) original imbalanced data, b) data generated using CBUS and c) data generated using SMOTE over-sampling.

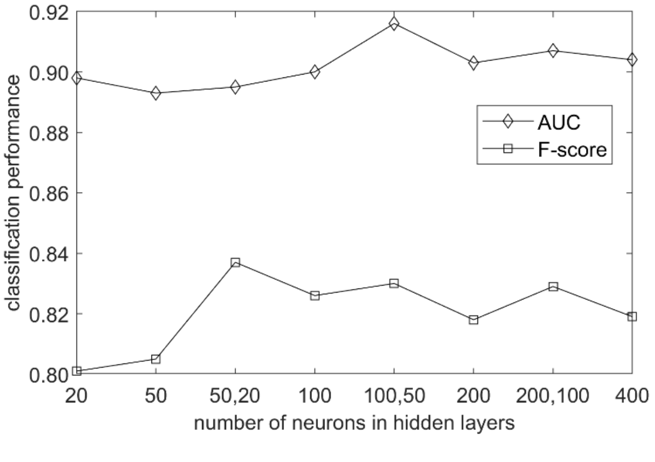


Fig. 8: Overfitting analysis of the neural network-based detection model

To demonstrate the effectiveness of the integrated fake detection model, its performance was compared with five state-of-the-art models used in previous studies for fake review detection:

- Sentiment analysis + NB (SA+NB) (Deng and Chen, 2014). Ratings were first used to categorize the restaurant reviews into positive (>3 stars) and negative classes (≤ 3 stars). Then, Harvard General Inquirer dictionaries[4] were compared with the texts of the restaurant reviews to obtain the list of positive and negative sentiment words. The words were also categorized into Place, Food, Travel and Quality. Then, NB was trained to detect fake reviews.
- Linear SVM (LSVM) using *n*-grams as linguistic features (Ahmed et al., 2018). It should be noted that different weighting schemes and *n*-gram sizes were tested by the authors of this method. In agreement with Ahmed et al. (2018), 50,000 unigrams were used with *tf.idf* weightings. The value of complexity parameter *C* was determined using a grid search in the range of $2^0, 2^1, \dots, 2^7$.
- Polarity-oriented unique attributes + LSVM (Polar.+LSVM) (Martinez-Torres and Toral, 2019). Following the authors of this method, the polarity-oriented unique attribute selection was performed by applying ANOVA to the set of unigrams with *tf.idf* weightings. Unique attributes were associated with four classes, namely (1) positive fake reviews, (2) negative fake reviews, (3) positive legitimate reviews, and (4) negative legitimate reviews. Martinez-Torres and Toral (2019) compared six different machine learning methods, showing that LSVM performed best among those methods with complexity parameter C=1.
- Word embeddings (Skip-Gram), *n*-gram and emotion representations + DFFNN (WE+emotion+DFFNN) (Hajek et al., 2020). Consistent with Hajek et al. (2020), the Skip-Gram model was trained for 100-dimensional embeddings with a context size of 5. Furthermore, the top 2,000 *n*-grams (unigrams, bigrams and trigrams) were selected according to their *tf.idf* weights. Finally, thirty lexicon-based emotion indicators were

[4] http://www.wjh.harvard.edu/~inquirer/homecat.htm

calculated to consider the sentiment-related information. DFFNN was trained using the mini-batch gradient descent algorithm with 100 and 50 neurons in the hidden layers.

- Psychological cues + LR (Psychol.+LR) (Li et al., 2020). Linguistic Inquiry and the Word Count 2015 (LIWC2015) dictionary was used to measure the psychological cues (affective, cognitive, perceptual, and social) in terms of the percentages of words matching those word categories. In addition, time distance, reviewer experience, review star rating and review length were considered in agreement with Li et al. (2020).

To overcome the problem of imbalanced classes, random under-sampling was employed for all the compared methods, which is consistent with earlier studies (Liu et al., 2019; Zhang et al., 2016; Budhi et al., 2021). Table 3 also shows AUC and F-measure performance of the baseline models reported above to demonstrate improvement obtained using the combination of behavioural and sentiment-dependent linguistic patterns. The proposed model consistently outperformed the compared fake review detection models for both evaluation measures, indicating good performance for both fake and legitimate classes. Overall, the proposed model significantly outperformed both the compared and baseline models. The results suggest that the models using linguistic features (either BoW or sentiment features) performed worse than those with behavioural features although sentiment (emotion) information improved the performance of the traditional BoW representation. This finding suggests that the linguistic features alone have limited capacity to detect fake reviews. Indeed, the model proposed by Li et al. (2020) and those proposed here using behavioural patterns significantly improved the classification performance (*p*=0.05, the Wilcoxon signed rank test) compared with their linguistic-based counterparts.

Table 3: Results of the experiments – classification performance

Model	AUC	F-score
SA+NB (Deng and Chen, 2014)	0.618±0.001	0.456±0.003
BoW+LSVM (Ahmed et al., 2018)	0.585±0.010	0.713±0.063
Polar.+LSVM (Martinez-Torres and Toral, 2019)	0.664±0.002	0.710±0.001

WE+emotion+DFFNN (Hajek et al., 2020)	0.744±0.002	0.732±0.003
Psychol.+LR (Li et al., 2020)	0.784±0.002	0.809±0.001
Baseline – sent-BoW features	0.714±0.001	0.696±0.006
Baseline – sent-Embeddings features	0.675±0.002	0.703±0.011
Baseline – behavioural features	0.890±0.001	0.799±0.004
The proposed model – all features	0.916±0.001*	0.830±0.001*

* significantly better performance (the Wilcoxon signed rank test at the *p*=0.05 level)

To check robustness of the results with respect to rating on reviews, the AUC values were calculated for each rating score. Fig. 9 shows that the proposed model performs well for all rating scores by consistently exceeding AUC of 0.9. Notably, fake reviews were correctly detected also for the extreme rating scores.

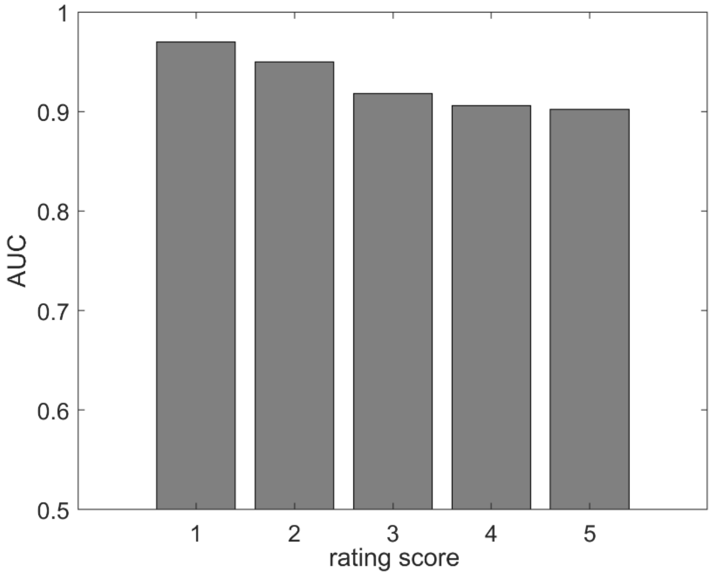


Fig. 9: Classification performance for different rating scores in terms of AUC

In the last run of experiments, the rating deviations r^{dev} of the models compared were calculated to evaluate the ability of the detection models to reduce the detrimental effect of fake reviews on average product rating (Table 4). Notably, the results demonstrate the major shortcoming of existing fake review detection methods. Namely, although existing methods identify fake reviews with relatively high accuracy (Table 3), they fail to provide a reliable rating estimate for the products. Table 4 presents the baseline evaluation given by the rating deviation obtained when no filtering is applied, i.e. no suspicious reviews are removed. A key problem of many of the existing automatic filtering methods is that many legitimate reviews are filtered out, which leads to significant distortion of average rating for individual products. As indicated in Table 4, such distortion effects are more serious for restaurants with lower average ratings. To prevent undesirable removal of ratings provided by legitimate reviewers, the ratings of fake reviews were removed only when a high confidence score was reported for the detection method. To determine the cut-off value for the confidence score, this study used the approach based on the ratio between the numbers of legitimate and fake reviews in the dataset and the ratio between their benefits and costs, respectively, as developed for imbalanced credit scoring data (Papouskova and Hajek, 2019). To estimate the ratio between the benefits and costs, relative importance of 7:1 was used, which was found to be most appropriate for spam detection (Zhang et al., 2014b). As a result, the confidence score required for disregarding a review rating was set to 0.93. Despite this setting, several methods (SA+NB, BoW+LSTM, and Polar.+LSVM) produced significant rating distortion (p=0.05, the Wilcoxon signed rank test), which can be attributed to a high error rate on the legitimate class. By contrast, Table 4 shows that the proposed model performed best with the mean value of rating deviation r^{dev}=0.1470 for all the restaurants included. For the compared methods, the baseline rating deviation was decreased only for Psychol.+LR (for good restaurants) and WE+emotion+DFFNN (for bad restaurants). Unlike the compared methods, the proposed model consistently decreased the average rating deviation for both restaurant categories (by 5% for good, by 3% for bad, and by 3.2% for all restaurants). The results in Table 4 also show that this decrease was statistically significant compared with the baseline scenario.

Table 4: Results of the experiments – average rating deviation

Model	r^{dev} (all)	r^{dev} (good)	r^{dev} (bad)
Baseline – no fake reviews discarded	0.1518	0.1011	0.1854
SA+NB (Deng and Chen, 2014)	0.3946	0.3130	0.4461
BoW+LSVM (Ahmed et al., 2018)	0.2737	0.2059	0.3172
Polar.+LSVM (Martinez-Torres and Toral, 2019)	0.2739	0.2060	0.3175
WE+emotion+DFFNN (Hajek et al., 2020)	0.1512	0.1011	0.1845

Psychol.+LR (Li et al., 2020)	0.1519	0.1004	0.1860
The proposed model	0.1470**	0.0976*	0.1797**

Note: all – all restaurants, good – restaurants with average rating avg(☆*r*) ≥ 4☆, bad – restaurants with average rating avg(☆*r*) < 4☆, ** significantly outperforming the baseline using the Wilcoxon signed rank test at p=0.05, * at p=0.10.

It should be noted that the labels of fake and legitimate reviews based on the Yelp's proprietary algorithm is only regarded as "near" ground truth, being prone to exaggerated false positive reviews (Luca and Zervas, 2016). To overcome the problem of the lack of truth labels, researchers have recently collected several datasets by generating convincing fake reviews. Therefore, to confirm the obtained results and address the issues of external validity, we checked the robustness of the proposed fake review detection model using a dataset of Indian restaurants[5] collected by Abri et al. (2020) and Gutierrez-Espinoza et al. (2020). This dataset consists of 110 reviews of three Indian restaurants, 50% of which are fake. The legitimate reviews were collected from online sources (Google), while the fake reviews were written by a group of university students. The credibility of legitimate reviews was ensured by selecting verified users only. The dataset was also balanced in terms of positive and negative reviews, but sentiment assignment was not available for this dataset. Therefore, we assigned the respective sentiment classes ourselves.

As a robustness check, we trained the proposed models on the dataset of Indian restaurants using the same experimental setup as for the Yelp dataset. Note that the data collection system and sampling process differed for the two datasets, which allowed us to verify the robustness of the fake review detection model. In agreement with Abri et al. (2020) and Gutierrez-Espinoza et al. (2020), we evaluated the performance of the model in terms of Accuracy and F-score (Table 5). In Table 5, we also present the results obtained by (1) Doc2Vec embeddings + decision tree (DT) / extreme gradient boosting trees (XGBoost) / multilayer perceptron (MLP) (Gutierrez-Espinoza et al., 2020) and (2) lexical features (linguistic cues such as

[5] https://github.com/asiamina/FakeReviews-RestaurantDataset

redundancy, pausality, average sentence length, and the number of adjectives) + random forest (RF) / MLP (Abri et al., 2020). Details on the hyperparameter settings can be found in the referenced studies. Note that due to the class-balanced dataset, there was no need to use the CBUS algorithm in this case. Despite this, we were able to increase the detection accuracy by more than two percent, which can be mainly attributed to the sentiment-dependent word embeddings.

Table 5: Classification performance on dataset of Indian restaurants

Model	Accuracy	F-score
Doc2Vec + DT (Gutierrez-Espinoza et al., 2020)	0.796	0.769
Doc2Vec + XGBoost (Gutierrez-Espinoza et al., 2020)	0.783	0.726
Doc2Vec + MLP (Gutierrez-Espinoza et al., 2020)	0.680	0.686
Lexical features + MLP (Abri et al., 2020)	0.791	0.770
Lexical features + RF (Abri et al., 2020)	0.755	0.737
Baseline – sent-BoW features	0.791	0.791
Baseline – sent-Embeddings features	0.800	0.800
The proposed model – all features	0.818	0.818

5. Discussion and Managerial Implications

This study confirms the superiority of machine learning approaches based on review- and reviewer-centric features (Wu et al., 2020). Compared to previous studies, the model developed herein has the advantage of being based on behavioural and sentiment-dependent linguistic features that effectively exploit the domain context. Indeed, it obtains better performance than the following five state-of-the-art models: SA+NB (Deng and Chen, 2014), BoW+LSVM (Ahmed et al., 2018), Polar.+LSVM (Martinez-Torres and Toral, 2019), WE+emotion+DFFNN (Hajek et al., 2020), and Psychol.+LR (Li et al., 2020).

In addition to this, some performance measures (Acc in particular or R^2) of other published models may be overestimated when using symmetrical samples (Soleymani et al., 2020). In reality however, there are not as many fake reviews as there are true reviews. The real performance of the models tested on symmetrical samples, such as those of Ahmed et al. (2018), Martinez-Torres and Toral (2019) and Hajek et al. (2020), is therefore lower.

The high detection performance of this model opens up the prospect of creating more efficient systems to detect fake reviews. Given its complexity, this kind of system could be set up by review platforms to certify reviews or to provide this service to other merchants via a label, for example. Indeed, the large number of fake reviews on sites such as TripAdvisor (The Times, 2018) shows that it is necessary to reduce them in order for this type of site to maintain its credibility and market power as a tourist guide. This is also important for e-commerce sites like Amazon to develop their sales.

Eliminating or reducing the visibility of fake reviews will also greatly diminish the benefits that their creator can derive from them (Gentina et al., 2020). As noted above, Yelp and other online platforms employ automated filtering algorithms to identify potentially fraudulent reviews. On the one hand, these reviews are not considered in the overall product rating score. On the other hand, the reviews remain visible to potential customers. A conservative approach is commonly used by the online platforms in practice, which in turn leads to placing legitimate reviews in the filtered category (Luca and Zervas, 2016). The present findings regarding the distorting effect of this approach on rating deviation contradicts this practice and suggests a different course of action. More precisely, a high confidence score for fake review filtering should be required to avoid the rating bias. The results of this study results show that the proposed detection model curtails this negative effect of fake reviews. Therefore, this research could be a useful aid for customers to find restaurants with more reliable rating scores (i.e., with less distortions induced by fake reviews). Honest restaurant managers benefit from this correction mechanism by enhancing consumer confidence (Moon et al., 2019). This may also have considerable managerial implications with respect to legal consequences of removing legitimate reviews. Indeed, posting a confidence score along with the review supports the transparency of online platforms and may prevent unjustified claims. Indeed, regulatory risks are associated with interfering with the online reviews that prompt actions by regulatory bodies.

Note that these risks include not only fake reviews but also incentivising positive reviews and moderating bad reviews. Therefore, rating reliability and consistency is important not only for consumers and businesses but also for authorities (Poddar et al., 2019). One possible application of the present model would be to set up alerts when the rating deviation measure exceeds a given value.

The proposed approach also has the potential to add new features to detect review manipulation and thus improve informativeness of the reviews in the hospitality industry. This, paper focused on sentiment-dependent linguistic patterns and demonstrated their effectiveness compared with existing approaches on a large real-world dataset. It showed that sentiment-related information should be considered when identifying linguistic features in suspicious reviews, indicating that the lack of such information makes it difficult for consumers, restaurant managers and online review platforms to detect fake reviews and make purchase decisions. For example, restaurant managers should implement systems to monitor the sentiment-dependent linguistic features in online review platforms and be proactive in reporting suspicious review activity.

6. Conclusion

Returning to the posed hypotheses, it is now possible to draw the following main conclusions: (1) Sentiment-dependent linguistic feature extraction is more effective for fake review detection. (2) Class imbalance must be considered in the process of training data selection. (3) To achieve a superior detection performance, linguistic features must be combined with behavioural features. (4) Balanced performance on both fake and legitimate class is needed to reduce the distortion effect of fake reviews on average product ratings.

A number of caveats must be noted regarding the present study. The most significant limitation lies in the fact that aspect level of granularity was not considered. By focusing on the review level, it was assumed that sentiment is consistent across aspects concerned in the review.

However, prior research suggests that the aspects associated with positive and negative fake and legitimate reviews may be different (Martinez-Torres and Toral, 2019). Further investigation and experimentation into aspect-based sentiment analysis is therefore strongly recommended. As such, it is recommended to add an aspect extraction component in future research. This research has thrown up many questions in need of further investigation. Further research should be done to establish whether alternative sentiment-dependent linguistic features are effective in fake review detection. Alternative sentiment-specific word embeddings and word weighting schemes should be investigated. Another promising future avenue of research might be to construct emotion-dependent linguistic features by considering additional sentiment features such as trust and uncertainty. Furthermore, when including spatiotemporal reviewer-centric features, the DFFNN model used should be replaced by a CNN model in future studies. The NN model can also be modified by implementing the rating deviation measure as its objective function to minimize the distortion effects of false review classification.

Acknowledgments

This article was supported by the scientific research project of the Czech Sciences Foundation Grant No. 19-15498S. We gratefully acknowledge the help provided by constructive comments of the anonymous referees.

References

Abri, F., Gutiérrez, L. F., Namin, A. S., Jones, K. S., Sears, D. R. (2020). Linguistic features for detecting fake reviews. In 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 352-359.

Ahmad, W., Sun, J. (2018). Modeling consumer distrust of online hotel reviews. International Journal of Hospitality Management, 71, 77-90.

Ahmed, H., Traore, I., Saad, S. (2018). Detecting opinion spams and fake news using text classification. Security and Privacy, 1(1), e9.

Barbado, R., Araque, O., Iglesias, C. A. (2019). A framework for fake review detection in online consumer electronics retailers. Information Processing & Management, 56(4), 1234-1244.

Barushka, A., Hajek, P. (2019). Review spam detection using word embeddings and deep neural networks. IFIP International Conference on Artificial Intelligence Applications and Innovations, Springer, Cham, 340-350.

Bravo-Marquez, F., Mendoza, M., Poblete, B. (2014). Meta-level sentiment models for big social data analysis. Knowledge-Based Systems, 69, 86-99.

Budhi, G. S., Chiong, R., Wang, Z. (2021). Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features. Multimedia Tools and Applications, 80(9), 13079-13097.

Cardoso, E. F., Silva, R. M., Almeida, T. A. (2018). Towards automatic filtering of fake reviews. Neurocomputing, 309, 106-116.

Chatzakou, D., Vakali, A. (2015). Harvesting opinions and emotions from social media textual resources. IEEE Internet Computing, 19(4), 46-50.

Chen, R. Y., Guo, J. Y., Deng, X. L. (2014). Detecting fake reviews of hype about restaurants by sentiment analysis. In International Conference on Web-Age Information, Springer, Cham, Management, 22-30.

Chen, L., Li, W., Chen, H., Geng, S. (2019). Detection of fake reviews: Analysis of sellers' manipulation behavior. Sustainability, 11(17), 4802.

Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. Journal of Big Data, 2(1), 1-24.

Deng, X., Chen, R. (2014). Sentiment analysis based online restaurants fake reviews hype detection. In: Han W., Huang Z., Hu C., Zhang H., Guo L. (eds) Web Technologies and Applications. APWeb 2014. Lecture Notes in Computer Science, 8710, Springer, Cham, 1-10.

Deng, Z. H., Luo, K. H., Yu, H. L. (2014). A study of supervised term weighting scheme for sentiment analysis. Expert Systems with Applications, 41(7), 3506-3513.

Dhaoui, C., Webster, C. M., Tan, L. P. (2017). Social media sentiment analysis: lexicon versus machine learning. Journal of Consumer Marketing, 34(6), 480-488.

Fang, Y., Wang, H., Zhao, L., Yu, F., Wang, C. (2020). Dynamic knowledge graph based fake-review detection. Applied Intelligence, 50(12), 4281-4295.

Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R. (2013). Exploiting burstiness in reviews for review spammer detection. In Proceedings of the International AAAI Conference on Web and Social Media, pp. 175-184.

Fusilier, D. H., Montes-y-Gómez, M., Rosso, P., Cabrera, R. G. (2015). Detecting positive and negative deceptive opinions using PU-learning. Information Processing & Management, 51(4), 433-443.

Gentina, E., Chen, R., & Yang, Z. (2020). Development of theory of mind on online social networks: Evidence from Facebook, Twitter, Instagram, and Snapchat. Journal of Business Research, 124, 652-666.

Gutierrez-Espinoza, L., Abri, F., Namin, A. S., Jones, K. S., Sears, D. R. (2020). Ensemble learning for detecting fake reviews. In 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), pp. 1320-1325.

Hajek, P., Barushka, A., Munk, M. (2020). Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining. Neural Computing and Applications, 32(23), 17259-17274.

Harris, C. G. (2012). Detecting deceptive opinion spam using human computation. Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence, 87-93.

Heydari, A., ali Tavakoli, M., Salim, N., Heydari, Z. (2015). Detection of review spam: A survey. Expert Systems with Applications, 42(7), 3634-3642.

Hou, Z., Cui, F., Meng, Y., Lian, T., Yu, C. (2019). Opinion mining from online travel reviews: A comparative analysis of Chinese major OTAs using semantic association analysis. Tourism Management, 74, 276–289.

Hussain, N., Mirza, H. T., Hussain, I., Iqbal, F., Memon, I. (2020). Spam review detection using the linguistic and spammer Behavioral methods. IEEE Access, 8, 53801-53816.

Jia, S. S. (2020). Motivation and satisfaction of Chinese and US tourists in restaurants: A cross-cultural text mining of online reviews. Tourism Management, 78, 104071.

Jindal, N., Liu, B. (2007). Analyzing and detecting review spam. Seventh IEEE International Conference on Data Mining (ICDM 2007), IEEE, 547-552.

Jindal, N., Liu, B., Lim, E. P. (2010). Finding unusual review patterns using unexpected rules. Proceedings of the 19th ACM international conference on Information and Knowledge Management, 1549-1552.

Kauffmann, E., Peral, J., Gil, D., Ferrández, A., Sellers, R., Mora, H. (2020). A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making. Industrial Marketing Management, 90, 523-537.

Kayser, V., Blind, K. (2017). Extending the knowledge base of foresight: The contribution of text mining. Technological Forecasting and Social Change, 116, 208-215.

Kennedy, S., Walsh, N., Sloka, K., Foster, J., McCarren, A. (2020). Fact or factitious? Contextualized opinion spam detection. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 344–350.

Kim, K., Park, O. J., Yun, S., Yun, H. (2017). What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management. Technological Forecasting and Social Change, 123, 362-369.

Kouloumpis, E., Wilson, T., Moore, J. (2011). Twitter sentiment analysis: The good the bad and the OMG! Proceedings of the International AAAI Conference on Web and Social Media, 538-541.

Lau, R. Y., Liao, S. Y., Kwok, R. C. W., Xu, K., Xia, Y., Li, Y. (2012). Text mining and probabilistic language modeling for online review spam detection. ACM Transactions on Management Information Systems (TMIS), 2(4), 1-30.

Li, F., Huang, M., Yang, Y., Zhu, X. (2011). Learning to identify review spam. Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, 11(3), 2488-2493.

Li, H., Chen, Z., Mukherjee, A., Liu, B., Shao, J. (2015). Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns. Proceedings of the International AAAI Conference on Web and Social Media, 634-637.

Li, J., Ott, M., Cardie, C., Hovy, E. (2014). Towards a general rule for identifying deceptive opinion spam. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 1, 1566-1576.

Li, L., Qin, B., Ren, W., Liu, T. (2017). Document representation and feature combination for deceptive spam review detection. Neurocomputing, 254, 33-41.

Li, L., Lee, K. Y., Lee, M., Yang, S. B. (2020). Unveiling the cloak of deviance: Linguistic cues for psychological processes in fake online reviews. International Journal of Hospitality Management, 87, May, 102468.

Lim, E. P., Nguyen, V. A., Jindal, N., Liu, B., Lauw, H. W. (2010). Detecting product review spammers using rating behaviors. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 939-948.

Liu, Y., Pang, B., Wang, X. (2019). Opinion spam detection by incorporating multimodal embedded representation into a probabilistic review graph. Neurocomputing, 366, 276-283.

Liu, Y., Pang, B. (2018). A unified framework for detecting author spamicity by modeling review deviation. Expert Systems with Applications, 112, 148-155.

Luca, M., Zervas, G. (2016). Fake it till you make it: Reputation, competition, and Yelp review fraud. Management Science, 62(12), 3412-3427.

Manaskasemsak, B., Tantisuwankul, J., Rungsawang, A. (2021). Fake review and reviewer detection through behavioral graph partitioning integrating deep neural network. Neural Computing and Applications, 1-14. doi: 10.1007/s00521-021-05948-1.

Martinez-Torres, M. R., Toral, S. L. (2019). A machine learning approach for the identification of the deceptive reviews in the hospitality sector using unique attributes and sentiment orientation. Tourism Management, 75, 393-403.

Mitra, S., Jenamani, M. (2021). Helpfulness of online consumer reviews: A multi-perspective approach. Information Processing & Management, 58(3), 102538.

Mohawesh, R., Tran, S., Ollington, R., Xu, S. (2021). Analysis of concept drift in fake reviews detection. Expert Systems with Applications, 169, 114318.

Moon, S., Kim, M. Y., Bergey, P. K. (2019). Estimating deception in consumer reviews based on extreme terms: Comparison analysis of open vs. closed hotel reservation platforms. Journal of Business Research, 102, 83-96.

Moon, S., Kim, M. Y., Iacobucci, D. (2021). Content analysis of fake consumer reviews by survey-based text categorization. International Journal of Research in Marketing, 38(2), 343-364.

Moraes, R., Valiati, J. F., Neto, W. P. G. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications, 40(2), 621-633.

Mukherjee, A., Venkataraman, V., Liu, B., Glance, N. (2013). What yelp fake review filter might be doing? Proceedings of the International AAAI Conference on Web and Social Media, 409-418.

Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., Ghosh, R. (2013b). Spotting opinion spammers using behavioral footprints. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 632-640.

Ofek, N., Rokach, L., Stern, R., Shabtai, A. (2017). Fast-CBUS: A fast clustering-based undersampling method for addressing the class imbalance problem. Neurocomputing, 243, 88-102.

Ott, M., Cardie, C., Hancock, J. T. (2013). Negative deceptive opinion spam. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 497-501.

Papouskova, M., Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning. Decision Support Systems, 118, 33-45.

Paul, H., Nikolaev, A. (2021). Fake review detection on online E-commerce platforms: a systematic literature review. Data Mining and Knowledge Discovery, 1-52. doi: 10.1007/s10618-021-00772-6

Pennington, J., Socher, R., Manning, C. D. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543.

Plotkina, D., Munzel, A., Pallud, J. (2020). Illusions of truth - Experimental insights into human and algorithmic detections of fake online reviews. Journal of Business Research, 109, 511-523.

Poddar, A., Banerjee, S., Sridhar, K. (2019). False advertising or slander? Using location based tweets to assess online rating-reliability. Journal of Business Research, 99, 390-397.

Rayana, S., Akoglu, L. (2015). Collective opinion spam detection: Bridging review networks and metadata. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August, pp. 985-994.

Ren, Y., Ji, D. (2017). Neural networks for deceptive opinion spam detection: An empirical study. Information Sciences, 385, 213-224.

Ruan, N., Deng, R., Su, C. (2020). GADM: Manual fake review detection for O2O commercial platforms. Computers & Security, 88, 101657.

Shan, G., Zhou, L., Zhang, D. (2021). From conflicts and confusion to doubts: Examining review inconsistency for fake review detection. Decision Support Systems, 144, 113513.

Shehnepoor, S., Salehi, M., Farahbakhsh, R., Crespi, N. (2017). NetSpam: A network-based spam detection framework for reviews in online social media. IEEE Transactions on Information Forensics and Security, 12(7), 1585-1595.

Shojaee, S., Murad, M. A. A., Azman, A. B., Sharef, N. M., Nadali, S. (2013). Detecting deceptive reviews using lexical and syntactic features. In Proceedings of the 2013 13th International Conference on Intelligent Systems Design and Applications, IEEE, pp. 53-58.

Schuckert, M., Liu, X., Law, R. (2016). Insights into suspicious online ratings: direct evidence from TripAdvisor. Asia Pacific Journal of Tourism Research, 21(3), 259-272.

Soleymani, R., Granger, E., Fumera, G. (2020). F-measure curves: A tool to visualize classifier performance under imbalance. Pattern Recognition, 100 (April), 107146.

Sun, C., Du, Q., Tian, G. (2016). Exploiting product related review features for fake review detection. Mathematical Problems in Engineering, Hindawi, 2016, 1-7.

Vidanagama, D. U., Silva, T. P., Karunananda, A. S. (2020). Deceptive consumer review detection: a survey. Artificial Intelligence Review, 53(2), 1323-1352.

Wang, J., Kan, H., Meng, F., Mu, Q., Shi, G., Xiao, X. (2020). Fake review detection based on multiple feature fusion and rolling collaborative training. IEEE Access, 8, 182625-182639.

Wu, Y., Ngai, E. W., Wu, P., Wu, C. (2020). Fake online reviews: Literature review, synthesis, and directions for future research. Decision Support Systems, 132, 113280.

Xiang, Z., Du, Q., Ma, Y., Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. Tourism Management, 58, 51-65.

Yen, S. J., Lee, Y. S. (2006). Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In Huang DS., Li K., Irwin G.W. (eds.) Intelligent Control and Automation. Lecture Notes in Control and Information Sciences, vol. 344, Springer, Berlin, pp. 731-740.

Yen, S. J., Lee, Y. S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications, 36(3), 5718-5727.

Yilmaz, C. M., Durahim, A. O. (2018). SPR2EP: a semi-supervised spam review detection framework. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, pp. 306-313.

Zeng, Z. Y., Lin, J. J., Chen, M. S., Chen, M. H., Lan, Y. Q., Liu, J. L. (2019). A review structure based ensemble model for deceptive review spam. Information, 10(7), 243.

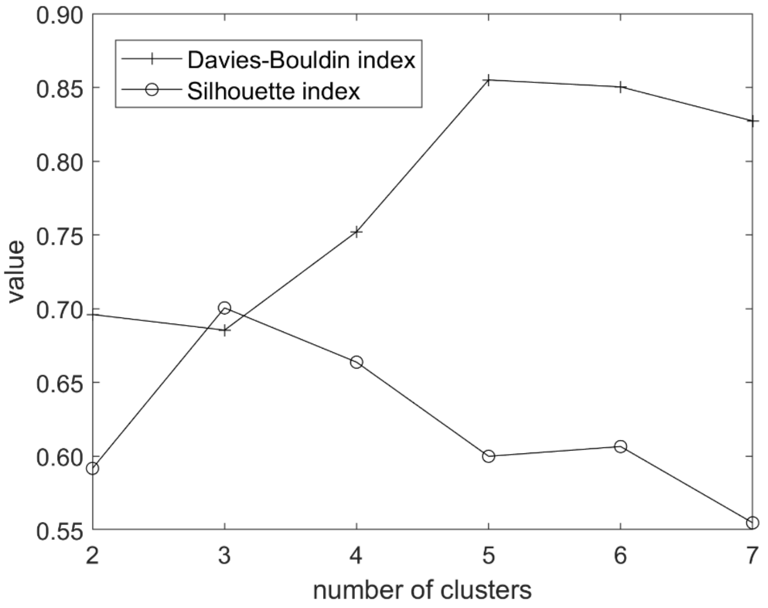
Zhang, K. Z., Zhao, S. J., Cheung, C. M., Lee, M. K. (2014). Examining the influence of online reviews on consumers' decision-making: A heuristic–systematic model. Decision Support Systems, 67, 78-89.

Zhang, Y., Wang, S., Phillips, P., Ji, G. (2014b). Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. Knowledge-Based Systems, 64, 22-31.

Zhang, D., Zhou, L., Kehoe, J. L., 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), 456-481.

Zhang, J. (2019). What's yours is mine: exploring customer voice on Airbnb using text-mining approaches. Journal of Consumer Marketing, 36(5), 655-665.

Appendix 1: Clustering quality indices


value
number of clusters
Davies-Bouldin index
Silhouette index