Detecting and Understanding Textual Deepfakes in Online Reviews

Peter Kowalczyk
University of Würzburg
peter.kowalczyk@uni-wuerzburg.de

Marco Röder
University of Würzburg
marco.roeder@uni-wuerzburg.de

Alexander Dürr
University of Würzburg
alexander.duerr@uni-wuerzburg.de

Frédéric Thiesse
University of Würzburg
frederic.thiesse@uni-wuerzburg.de

Abstract

Deepfakes endanger business and society. Regarding fraudulent texts created with deep learning techniques, this may become particularly evident for online reviews. Here, customers naturally rely on truthful information about a product or service to adequately evaluate its worthiness. However, in light of the proliferation of deepfakes, customers may increasingly harbour distrust and thereby affect a retailer's business. To counteract this, we propose a novel IT artifact capable of detecting textual deepfakes to then explain their peculiarities by using explainable artificial intelligence. Finally, we demonstrate the utility of such explanations for the case of online reviews in e-commerce.

1. Introduction

Deepfakes pose severe threats to business and society. Deepfakes can be regarded as fakes generated by deep learning (DL) technologies [1, 2, 3]. No matter if the respective DL techniques are intentionally abused to spread misinformation or explored out of curiosity, the deepfakes created can entail serious consequences.

A warning example was given by the out-of-character deepfake video of former U.S. President Barack Obama calling Donald Trump a "total and complete dipshit" [4, 5]. The U.S. House Intelligence Committee recognized the potential threats caused by deepfakes ahead of the 2020 U.S. presidential election and therefore carried out an extensive investigation to prepare adequate counteracts [6]. Another case from 2019 illustrated the increasing relevance of deepfakes for companies, when a fraudster used artificial intelligence to mimic the voice of a manager's superior and scam $243,000 [7].

The aforementioned examples are just the tip of the iceberg. In light of rapid innovation in the field of DL, the risk posed by deepfakes is only reinforced. The more refined the underlying algorithms the more real and thus believable the fictitious output [2]. Thus, it becomes ever more challenging for humans to distinguish deepfakes from reality. In addition, due to the increasing ease of use and ipso facto, the ongoing democratization of DL, the proliferation of the technologies to generate deepfakes seems inevitable [2].

Besides the rather prominent image-based deepfake examples, fraudulent texts may also cause great harm [8]. Machine generated, deceptively real texts can spread rapidly due to a worldwide highly connected information network and thus may heavily affect people and mislead decision-makers. In particular, this becomes evident for online reviews. Here, customers naturally rely on the opinions of previous buyers to evaluate whether a purchase is reasonable [9]. Hence, these reviews can directly affect a company's reputation and profitability [10, 11, 12]. Faced with misinformation in the form of fake reviews at great scale consumers can no longer take anything for granted. Consequently, they are left to decide whether it is worth taking the gamble and committing to the purchase or refraining from it entirely. This uncertainty poses a great financial risk to honest retailers and service providers—especially the smaller ones who may be more dependent on authentic and positive reviews. In the worst case, fake reviews can result in long-term brand erosion and may ultimately lead to a downward spiral of distrust through word-of-mouth and reduced sales.

To overcome these novel challenges to online commerce, reliably identifying deepfakes in reviews is crucial. Hence, an automatic and fast detection of textual deepfakes at a great scale is the go-to target. To this end, the deployment of an information system offers a promising solution. More specifically, predictive analytics could be employed to validly detect deepfakes. Predictive analytics commonly refers to data-driven machine learning (ML) models that aim to make predictions about previously unseen instances [13, 14]. However, as deepfakes may gradually change to circumvent a conventional detection system, understanding the nature of such fakes could

URI: https://hdl.handle.net/10125/79516
978-0-9981331-5-7

prove very useful. Explainable artificial intelligence (XAI) provides insights into an algorithm's decisions. Such insights can in turn be used to explore the peculiarities of artificially generated online reviews and thus play a major role in automated deepfake identification. However, concerning fake online reviews in particular—to the best of our knowledge—no satisfying solution has been presented to date.

Against this backdrop, we propose an IT artifact that acts as a detector and insight generator for the case of deepfake online reviews by leveraging the potentials of DL and XAI. For this purpose, we provide a background on synthetic text generation with DL, text fraud detection and XAI. Subsequently, as we choose to follow a design-oriented research approach to develop an IT artifact, we adapt our study to the guidelines as proposed by Hevner et al. [15]. The next section is concerned with the detailed description of the artifact to derive valuable knowledge for deepfake detection in online reviews. The remainder of the paper focuses on the deduction of these insights for the specific use case of online reviews which opens up avenues for further research.

2. Theoretical Background

2.1. Text Generation with Deep Learning

DL is a vivid and constantly evolving field of research. This holds, among others, for DL-based synthetic text generation. In the following, we consider two major approaches to DL-driven synthetic text generation.

Generative Adversarial Networks (GANs) can take a random, unstructured input (i.e., Gaussian noise) and apply the adequate transformations such that entirely new but strikingly realistic data entities are created [16]. To this end, two deep neural networks (i.e., the discriminator and the generator) are used to try to outsmart each other by competing in a zero-sum game. Whereas the discriminator is capable of classifying either an original or a fake sample as such, the generator uses these classifications to gradually create more realistic data instances. After sufficient training, the generator ideally creates fake texts that are indistinguishable from real ones. GANs especially stand out due to the fact that they perform well on unstructured data (e.g., textual data). On the other hand, GANs can be difficult to fine-tune due to heavy parameter oscillations [17]. In addition, the *mode collapse phenomenon* can occur, that is, the generators outputs gradually become less diverse due to over-optimizing for a particular discriminator feedback [17]. Regarding GAN-based text generation popular approaches include *RelGAN* [18], *LeakGAN* [19] and *SeqGAN* [20].

Generative Pretrained Transformers (GPTs) are another popular type of text generation method. Contrary to previous architectures employed for text generation, like for example recurrent neural networks, the transformer architecture protects the model against suffering from short-term memory, meaning that these models can select and retain the relevant information for efficient text generation for as long as it is needed—assuming infinite computing power hypothetically. To be more precise, they feature the so-called attention mechanism as introduced by Vaswani et al. [21] that in contrast to previous approaches shifts the focus to the necessary bits of information of the whole context rather than relying on a finite reference window. Whereas *generative* in GPT refers to the usage context of the architecture, that is, to produce a new text, the *pretrained* relates to the fact that the model is already trained on a big linguistic corpus and therefore has a fundamentally good understanding of natural language [21].

Language models like GPT and GPT-2 are considered superior both in terms of the quality and diversity of the produced text compared to GAN-based approaches [22, 23, 24]. Moreover, as sophisticated transformer models pretrained on large textual corpora are already available and thus just require a fine-tuning for the specific text generation task, this also comes as a great benefit compared to GANs which often require starting from scratch with bad initial performance.

2.2. Automated Text Fraud Detection

Detecting fraud in online reviews is a perennial topic in research [10, 11]. This section briefly outlines some of the recent advances in deception detection for the case of online reviews by boiling down selected articles from extant literature to the utilized ML models as well as feature sets. In addition, we explore current articles on textual deepfake detection in particular.

Among the more popular ML approaches to fake detection in online reviews are supervised methods such as *Logistic Regression (LR)*, *Support Vector Machines (SVMs)*, *Random Forests (RFs)*, *Extreme Gradient Boosting (XGB)* and *Artificial Neural Networks (ANNs)* [25, 26, 12]. Regarding the features two types can broadly be distinguished—namely, content-based and linguistic [27, 25]. Whereas the former type of features originates from the review content [25, 26, 28, 10], the latter refers to language-based peculiarities of the texts [25, 29, 30, 28, 10].

With respect to textual deepfake detection, few

approaches have been presented to date. To identify deepfake tweets created with GPT-2 Fagni et al. [31] use LR, SVM and RF. Besides, Zhong et al. [32] and Zellers et al. [33] work on the detection of fake news through ANNs.

In light of the extant literature on fake online review detection, we found a variety of approaches. However, only few articles specifically deal with the detection of artificially created textual deepfakes. Moreover, none of these articles investigates the case of deepfake online reviews. Lastly, the explanation of deepfake detectors remains unconsidered which puts significant emphasis on this research direction by making the decisions of deepfake detection systems transparent and comprehensible to researchers and practitioners.

2.3. Explainable Artificial Intelligence

Predictive analytics is traditionally prone to the trade-off between accuracy and interpretability, that is, the more advanced the approaches to predictive analytics—i.e., ensemble or DL models—the better the results but the harder to interpret the corresponding models [13, 14, 34]. Whereas interpretable models are commonly referred to as white boxes, their usually better performing contenders are frequently regarded as black boxes [35]. XAI denotes a set of methods to interpret models that have long been considered as black boxes [36, 37, 38]. More specifically, XAI enables the creation of adequate explanations for black boxes and thereby provides multiple benefits (e.g., mitigate risks associated with bias, compliance, security or facilitate performance monitoring) [39, 37].

There are multiple XAI approaches with their respective benefits and drawbacks. In the context of this study, we favor SHAP over the other approaches as it is a model-agnostic approach (i.e., it is applicable to any black box model) that unifies a class of multiple existing interpretation methods [34]. In addition, SHAP employs additive feature attribution to induce simplicity. This is done by summing up the effects of all the features measured through so-called Shapley values to approximate the output of the original model [34]. Shapley values originate from cooperative game theory and measure the importance of each player to the overall effect as well as their respective expectable pay-off [40]. This concept is transferred to predictive analytics models with SHAP to determine the individual contribution of every feature per prediction and thus provide both local as well as global explanations [34]. Besides, compared with other current methods SHAP is advertised for its superior computational performance, capability to capture non-linearities and overall better consistency with human intuition [41, 34]. As a detailed description of SHAP and especially the underlying math is beyond the scope of this paper we refer the interested reader to the article by Lundberg et al. [34].

3. Design-oriented Research

To explore the nature of fake online reviews and thus provide utility to practitioners, we pursue a design-oriented research approach and design an IT artifact. Through this we want to enable transparency and encourage transferability [42]. To this end, we follow the guidelines as proposed by Hevner et al. [15]:

- **Problem Relevance**: Detecting fake online reviews is critical to customers and retailers and is getting more and more difficult due to rapid innovation in the field of synthetic text generation. To cope with this trend, it is essential to examine the characteristics of deepfakes by highlighting prevalent differences compared to genuine reviews.
- **Research Rigor**: For the purpose of our research, we deploy well-established state-of-the-art DL and XAI techniques.
- **Design as a Search Process**: To better understand the nature of deepfake reviews we set up the research project as a search process, that is, we propose an effective IT artifact but at the same time give room for further development.
- **Design as an Artifact**: We design an IT artifact that consists of three components to ultimately extract insights to the nature of deepfakes.
- **Design Evaluation**: To assess the effectiveness of the IT artifact we leverage preexisting data comprising almost 30,000 unique and genuine online reviews to first generate and subsequently detect deepfakes. Furthermore, we employ XAI to gather insights to the developed detector.
- **Research Contribution**: We sketch out an IT artifact that allows us to explore the nature of deepfakes in the context of online reviews through XAI. In light of the emerging threats resulting from text fraud, we state the importance to investigate the various scenarios in advance for the purpose of an early and effective anticipation to prevent or at least reduce possible harm associated.
- **Research Communication**: The proposed IT artifact enables researchers and practitioners

to conduct an analysis of the peculiarities of machine-forged online reviews and therefore helps to mitigate the emerging threat posed by textual deepfakes.

4. Artifact Description

The proposed artifact (cf. Figure 1) consists of three components. The first component is fed with original reviews to train the deepfake generator for the specific use context (cf. subsection 4.1). Given the full corpus with a perfectly balanced amount of fictitious and real online reviews a classifier is trained in the consecutive step to identify the fake texts (cf. subsection 4.2). Finally, the third component of the artifact (cf. subsection 4.3) aims to derive valuable insights into the nature of the deepfakes to later assist the generation of appropriate practical guidance which in turn supports the development of adequate countermeasures.

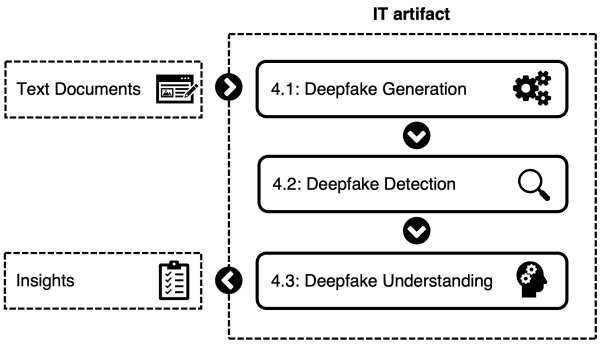

Figure 1. Architecture of the IT artifact

4.1. Deepfake Generation

The first component of the presented artifact deals with the creation of artificial texts. Following the narrative of the theoretical background on language generation models and the claimed superior performance of the transformer models for text generation we decide to opt for GPT-2. To that aim, the transformer model at hand is fine-tuned for the specific input (i.e., the online reviews) to then produce sophisticated fake texts.

4.2. Deepfake Detection

The second component is concerned with the detection of deepfakes. For this purpose, we (i) employ standard natural language preprocessing techniques to then (ii) derive linguistic as well as content-based features. This is required to facilitate the subsequent task with an adequate decision basis—i.e., the feature set. Given such feature set, we leverage ML to (iii) perform a classification task and predict the fraudulent reviews. The details of these steps are outlined below.

A common proceed prior to feature engineering is to preprocess through standard text mining techniques. Here, typical methods include—but are not limited to—lower-casing, stop word and punctuation removal, tokenization as well as stemming and lemmatizing [43, 44]. This procedure does not affect the meaning of the texts, but rather assists with feature assessment and saving computing capacities.

To derive the linguistic features in order to facilitate the subsequent classification task, we employ lower-casing to then adapt the nine linguistic-based cues for deception detection put forward by Zhou et al. [45]—namely, *quantity*, *diversity*, *complexity*, *specificity*, *expressivity*, *informality*, *affect*, *uncertainty*, and *non-immediacy*. The rationale behind this choice is twofold. First, these constructs are based on various theories on deception and second they have proven themselves to be highly suited to derive a feature set for text-based fraud detection [27, 25, 46, 47]. With respect to the linguistic constructs the corresponding features are as follows [45]:

- **Quantity**: Fraud reviews typically tend to contain fewer words than original ones. Therefore, we suggest to use the following self-explanatory features: *WordCount*, *VerbCount*, *NounCount*, *AdjectiveCount*, *AdverbCount*, *PronounCount*, *SentenceCount*, *PunctuationCount*
- **Diversity**: As deceptive reviews are perceived to be limited in terms of vocabulary usage we propose the feature *LexicalDiversity* to capture the ratio between the number of unique words to the total amount of words per text. In addition, we access significance through stop word associated features *StopwordCount* and *NotStopwordCount* as well as their respective ratios *StopwordRatio* and *NotStopwordRatio*—again compared to the overall number of words respectively—and the self-explanatory feature *StopwordToNotStopwordRatio*.
- **Complexity**: The higher the linguistic complexity of a text, the less likely it is a fraud. To measure the complexity degree we determine the average word length (*AvgWordLength*) and words per sentence (*AvgCharactersInSentence*) as well as the average ratio of punctuations per sentence (*Pausality*).
- **Specificity**: Fake texts are assumed to feature different sensorial perceptions (e.g., sounds,

smells, physical and visual sensations). Therefore, we compute the ratio of words in the context of perceptions compared to the total number of words (*PerceptionRatio*).

- **Expressivity**: Fake texts are perceived to predominantly include words that indicate expressions such as adjectives and adverbs to falsely appear to be genuine. Therefore, *Emotiveness* corresponds to the ratio between adjectives and adverbs to nouns and verbs.

- **Informality**: The feature *TypoRatio* indicates the number of misspelled words compared to the total number of words in a review text. The common belief is that fraudulent texts contain more mistakes than genuine texts.

- **Affect**: Affective metrics give indications of the current emotional state and may therefore be very helpful to distinguish fake reviews from real ones. Hence, we determine the portion of words with positive connotation in relation to the total number of words (*PositiveAffect*) et vice versa for the words with rather negative overtones (*NegativeAffect*). In addition, we access the *AffectRatio* (i.e., the number of words with affection to the total amount of words) and *PleasRatio* which indicates the ratio between the words of pleasantry (e.g., delightful, spirit, indulgence or comfortable) to all words.

- **Uncertainty**: This linguistic cue refers to contradictory stylistic figures in the case of text fraud and can for example be measured by the ratio of modal verbs to the total number of words (*ModalVerbRatio*) or the ratio between number of words associated with uncertainty (e.g., yet, careful, hesitant, tendency, hit, undefined) to the total amount of words in the texts (*UncertainRatio*).

- **Non-immediacy**: The last of the nine constructs assumes fraudulent texts to be written with submissive language due to feelings of guilt. Hence the variables *IndividualCount*, *GroupCount* and *SelfCount* each assess the amount of words connected to an individual, a group or a first-person speaker respectively. In addition, the variables *IndividualRatio*, *GroupRatio* and *SelfRatio* accordingly put these counts in proportion to the total number of words.

Beside these linguistic constructs we also suggest to extract content-based features. For this purpose, topic models—e.g., extracted with the well-established Latent Dirichlet Allocation (LDA) as proposed by Blei et al. [48]—might be appropriate. LDA captures the associations between words by examining their use context to create corresponding topics [48]. These topics can in turn be attributed to the texts according to their calculated appearance, that is, the higher the presence of a specific topic, the higher its respective feature value. As the computational effort associated with LDA can be considerably high for the original texts, we apply all of the introduced preprocessing techniques [49].

Subsequently, after obtaining the final set of features the actual classification task can be performed to detect the deepfake reviews. Here, we recommend comparing the performance of multiple ML models for a hold-out test set regarding the evaluation metrics and thus select the superior model with the best set of hyperparameters over the others [50, 13].

4.3. Deepfake Understanding

Having determined the best performing classification model to identify fake reviews based on the feature set, the IT artifact's next component can be employed. As the selected superior model presumably belongs to the rather complex end of the spectrum of ML models and hence can be regarded as a typical black box from an interpretability perspective, adequate XAI tools such as SHAP are required to gather valuable insights and develop an understanding to the algorithm's decisions. It should be noted that this understanding is heavily dependent on the prior selection of features and we therefore would like to stress the importance of a well-thought feature set. Ultimately, practitioners concerned with the global and local explanations of XAI must agree on the salient anomalies of deepfake reviews in order to assemble the most relevant insights that represent the outcome of the IT artifact.

5. Demonstration and Evaluation

To assess the performance and therefore the overall utility of the proposed artifact to validly detect deepfakes and assist the deduction of knowledge, we consider a data set with online reviews of Walmart Inc. retrieved from the Kaggle platform with around $30,000$ unique reviews[1]. Since we are only interested in the review texts we omit the other columns provided with the data set. Furthermore, we delete instances without a textual description to finally receive $24,617$ individual

[1] https://www.kaggle.com/promptcloud/walmart-product-reviews-dataset

review texts. In addition, we verify the authenticity of the reviews by performing plausibility checks with an extensive exploratory data analysis. This is done by three independent experts who randomly select five percent of the articles to then check them in terms of content-relatedness (e.g., the review is associated with the right product), product rating and verification label. Hereby we ensure that the deepfake generator learns from real review texts.

To train the deepfake generator, we use the pretrained 124M implementation of GPT-2 in the Python programming language[2] with a batch size of eight and ten million steps. Next, we double our database by drawing deepfake online reviews from the fine-tuned language generation model and thus obtain a perfectly balanced and labeled data set.

Prior to deepfake detection in the second step of the proposed artifact, we shuffle the data at random and apply the text mining techniques as well as feature deduction steps as outlined in section 4.2. Here, we use the Python package NLTK[3] to preprocess the texts for the derivation of the content-based features with another package (i.e., Gensim[4]). In addition, we employ the linguistic dictionary by General Inquirer[5] to assess the linguistic features. Regarding LDA we conclude that 13 seems to be the ideal number of topics as the coherence score is the highest here with 0.575. The topics are briefly listed in Table 1 along with their four most important terms according to the calculated probabilities and our chosen designation:

Table 1. Retrieved Topic Models.

	Descriptive Terms	Designation
1	tv, roku, remote, unit	streaming media
2	watch, get, like, would	interest
3	player, dvd, cable, channel	connection
4	buy, month, ago, purchase	purchase
5	phone, get, work, samsung	work
6	tv, picture, great, quality	visual media
7	review, part, promotion, collect	advertisement
8	great, product, price, good	deal
9	easy, use, set, great	installation
10	buy, love, son, gift	gift
11	screen, work, get, ipad	portable device
12	sound, good, quality, speaker	audio & hifi
13	camera, home, house, system	smart home

Given the complete feature set for the reviews, we train three ML models on the training set—namely, a dummy classifier (DC), a RF and XGB—to identify the deepfakes. In addition, we perform a five fold cross-validation and hyperparameter tuning via a grid search to then evaluate the tuned model's performance on the hold-out test set. To access the performance of the classifiers we employ oft-cited metrics such as accuracy (i.e., the proportion of correct predictions among the total number of samples) since our data set is highly balanced, precision (i.e., the fraction of true positive predictions among the positive predicted), recall (i.e., the fraction of relevant samples that were retrieved) and the harmonic mean of precision and recall, i.e., the F_1-Score [51].

These measures can be obtained from the confusion matrix which contrasts the predicted with the actual class. Resulting from these two classes the matrix in binary classification consists of four quadrants—namely, true positive (TP), false negative (FN), false positive (FP), and finally true negative (TN). Accordingly, the utilized metrics are computed as follows [51]:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F}_1 = 2 \cdot \frac{\text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}$$

When reviewing the results for the metrics rounded to five decimals, it stands out that XGB performed best regarding the holdout set (cf. Table 2) and is therefore selected for the next stage of the artifact.

Table 2. Comparison of the Deepfake Detectors.

Model	Accuracy	Precision	Recall	F_1-Score
DC	0.50696	0.50321	0.51299	0.50806
RF	0.88017	0.86594	0.89748	0.88143
XGB	**0.88697**	**0.87307**	**0.90362**	**0.88808**

To gain an understanding to the decisions of the chosen XGB model we rely on global as well as local explanations with SHAP. In essence, the higher the SHAP value the more likely the model predicts a deepfake et vice versa for low values.

For the purpose of global understanding, we look at the bee swarm plot (cf. Figure 2) provided with the SHAP implementation by [34]. Here, the 20

[2] https://github.com/minimaxir/aitextgen

[3] https://www.nltk.org

[4] https://radimrehurek.com/gensim/

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

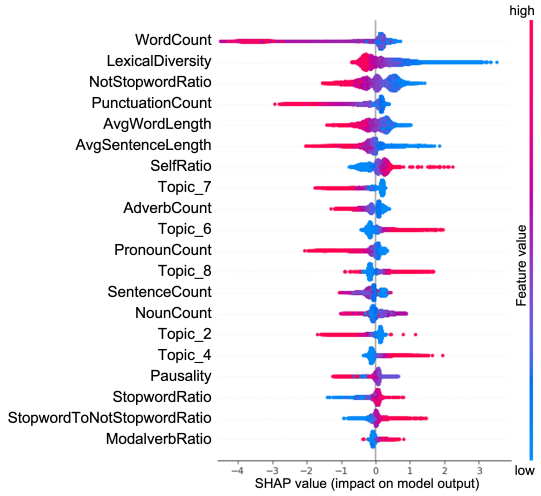

Figure 2. Global Understanding via SHAP Analysis.

most important features for the prediction are listed in descending order. The remaining features will not be discussed due to their comparatively minor importance.

The utmost important feature is *WordCount*—i.e., the higher the number of words in a review, the less likely it is a fake. Figure 2 illustrates that some reviews show particularly high feature values and thus consist of many words with a clear indication to the absence of a fraud. As for the second important feature—*LexicalDiversity*—a similar impression emerges, that is, the less unique the used words in a review compared to the number of total words, the more likely it is classified as a fraud [52, 45]. This is in line with the conjecture of a review being rich in lexical diversity [53]. Based on the algorithm's decisions *NotStopwordRatio* implies that the deepfake review relies more on stop words than an original online review. In accordance with the most important feature, the longer the review text, i.e., the more punctuation is employed (*PunctuationCount*), the higher the probability of the review being authentic. Similarly, the longer the words on average (*AvgWordLength*), the lower the SHAP value and thus the probability of predicting a text of being fraudulent.

The remaining features can be analyzed in a similar fashion. However, since referring to every feature in detail is beyond the scope of this paper, we rather shift our attention towards a few notable examples. As for the feature *SelfRatio*, the impact on the SHAP value is inverted. This indicates that in contrast to our initial assumption, the more prominent self-relating terms, the higher the SHAP value and thus the more likely a sample is classified as a deepfake. Similarly, this inverted impact can also be observed for the content-based feature *Topic_6*. As shown before, *Topic_6* mainly consists of contextual information about visual media. So, it seems fair to say that in our case such information increases the likelihood that the review is a fake. Lastly, we attribute *Pausality* to be kind of a mixed bag, which means that there is no clear indication of the feature value impact on the SHAP value. Solely the very high feature values (e.g., a high degree of punctuation within the sentences on average) suggest low SHAP values and thus lower the probability of predicting a fake online review.

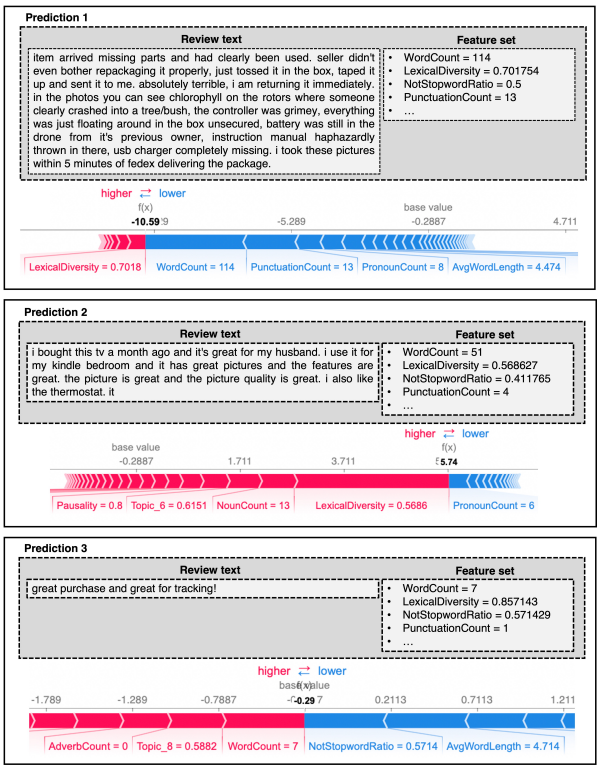

Figure 3. Local Understanding via SHAP Analysis.

Regarding the local explanations for each prediction, Figure 3 depicts the respective review text and a snippet of the corresponding feature set along with the explanation as per the so-called force plot. Here, the base SHAP value indicates the prediction of the XGB model without any features given for a text from the test set. However, since all features are available for the prediction, the complete feature set is considered. Hence, the plot highlights the main drivers for the specific prediction outcome according to the value and importance of the respective features. We reflect upon three contrasting scenarios in terms of the SHAP values—one with a clear indication of considering an authentic review (Prediction 1), the second with the opposite prediction of dealing with a fraud (Prediction

2) and the last being relatively indecisive (Prediction 3). We note that according to the correct labels the prediction holds true for the first two cases, whereas the last rather indecisive example is a generated deepfake.

As for the first prediction, the comparatively long text and thus the rather high variables *WordCount* and *PunctuationCount* heavily shift the resulting SHAP value downwards from the base value to indicate a genuine online review. Interestingly, contrary to the impression through the global explanation the features *VerbCount* and *LexicalDiversity* are part of the opposing effect. However, as the underlying SHAP values for these two features are not as high, this opposite flow does not significantly affect the prediction and thus turn around the outcome.

With regard to the second local explanation for the prediction of a rather obvious fake review the feature values for *LexicalDiversity*, *NounCount*, *Topic_6* and *Pausality* are considered to be the main decision drivers for the algorithm's choice. Notably, the review contains more stop words compared to the first analyzed prediction of the real text. This is in line with the overall effects of the features on the SHAP values as discussed previously in the context of the global explanation.

The last example investigated is relatively short and thus may prove to be a problem for an accurate classification. This idea is substantiated by the rather indecisive resulting SHAP value near the base value. Although to a practitioner this review might seem reasonably genuine, it is a deepfake. Here, the rather low SHAP values compared to the other two cases do not provide sufficient nor valuable explanations for the feature impact on the prediction. Thus, it might be beneficial to further investigate the specific case.

The above analysis for the developed IT artifact provides practitioners with valuable insights regarding the detection of deepfakes. In more detail, the following guidelines can be derived for the specific use case of online reviews:

- **Length**: Generally speaking, the longer the review, the words or the sentences, the less likely it is a forgery. Thus, a minimum length is recommended to demand.
- **Diversity**: A practitioner should carefully look at the linguistic diversity of the text which might indicate a fraud if it is rather unilateral.
- **Abundancy**: Genuine reviews are rather rich in information compared to their fictitious counterparts.
- **Topic reference**: Depending on the context, certain predominant themes may highlight the presence or absence of a fake and therefore should be taken into account.
- **Case-specificity**: If something seems peculiar or especially in the case of a rather indecisive prediction, a practitioner should carefully investigate the specific incident. Thus, despite the benefits of a fully automatic detection system manual checks might be required in some cases.

6. Discussion and Conclusion

This paper sheds light on the detection of the emerging threat posed by deepfakes for the specific case of online reviews. For this purpose, it provides a background on the generation and detection of such textual deepfakes as well as XAI as a toolbox to provide explanations to the algorithm's decisions. Based on this background, a corresponding research gap for online review deepfake detection and understanding is revealed. To contribute to closing this gap, we opt for a design-oriented research approach to develop an IT artifact which provides utility to researchers and practitioners. To assess its utility, we conduct an evaluation for the case of an online review data set. Related to this specific case, the results indicate a high success rate for the automatic detection by means of XGB and the chosen feature set (e.g., F_1=0.88808). In addition, XAI enables further investigation to develop a better understanding of the algorithm's decisions which turns out to be particularly useful for unclear predictions. Regarding the conducted demonstration of the IT artifact, we note that the results are based on the specific implementation, that is, the data set and thus selected language, language model, feature set, ML models, and choice of the XAI tool as well as the interpretation itself. Accordingly, the results may differ for another setup.

Nevertheless, this research holds several valuable implications—both from a theoretical as well as practical perspective. First and foremost, it presents a novel IT artifact consisting of three components (i.e., generation, detection and explanation) to encounter textual deepfakes in advance, which is both highly scalable and generalizable [42]. Hence, researchers and practitioners can easily transfer this conceptual IT artifact to another domain (e.g., financial reports or news texts) or likewise explore the effects for other types of generation models, detection algorithms and XAI methods. Regarding the practical implications, we recognize a high utility to the developed IT artifact due to (i) the proven high success rate in automatically detecting a large amount of deepfakes. Moreover, the system (ii) enables an automatic large-scale exclusion

of fraudulent texts and (iii) highlights the unclear edge cases while (iv) providing additional insights for further investigating these. Thus, with respect to the specific case of online reviews retailers and customers are supported with a fast, effective and interpretable tool which helps to maintain or increase trust in online commerce.

In conclusion, the paper opens opportunities for further research to assist the detection and explanation of deepfakes. Thus, it might be reasonable to explore other areas that may currently or in the future be affected by textual deepfakes. Regarding the evaluation, other transformer models, languages, data sets, ML models, feature sets, evaluation metrics or XAI methods could be employed and compared. Moreover, as the evaluation suggests, it might be beneficial to work on the automatic textual deepfake detection for the rather short texts which may provide a too small and thus ambiguous basis for an algorithm's decision-making. Besides the exclusive use of texts, incorporating further meta information such as the rating, review pictures or publication date among others might be rich in potential for future research. Furthermore, such a conceptual tritone (i.e., the IT artifact) can be adapted to various domains apart from online reviews. In addition, an automatic tool could be developed (e.g., in the form of a browser add-in) to indicate the probability of a text being fraudulent and thus yield great benefits to business and society. However, this tool should at best be aware of both—fakes generated by humans and machines in parallel. Lastly and generally speaking, as the derivation of appropriate guidelines depends on the people involved, we also emphasize the importance of human actors within the process to remedy the rising threat posed by deepfakes.

References

[1] M. Westerlund, "The emergence of deepfake technology: A review," *Technology Innovation Management Review*, vol. 9, pp. 39–52, 2019.

[2] J. Kietzmann, L. W. Lee, I. P. McCarthy, and T. C. Kietzmann, "Deepfakes: Trick or treat?," *Business Horizons*, vol. 63, no. 2, pp. 135–146, 2020.

[3] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *CoRR*, vol. abs/2004.11138, 2020.

[4] C. Vaccari and A. Chadwick, "Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news," *Social Media + Society*, vol. 6, no. 1, pp. 1–13, 2020.

[5] D. Mack, "This psa about fake news from barack obama is not what it appears," *BuzzFeed News*, May 2019.

[6] D. O'Sullivan, "Congress to investigate deepfakes as doctored pelosi video causes stir," *CNN*, Jun 2019.

[7] J. Damiani, "A voice deepfake was used to scam a CEO out of $243,000," *Forbes*, Sep 2019.

[8] R. DiResta, "Ai-generated text is the scariest deepfake of all," *Wired*, Jul 2020.

[9] X. Li and L. M. Hitt, "Price effects in online product reviews: An analytical model and empirical analysis," *MIS Quarterly*, vol. 34, no. 4, pp. 809–831, 2010.

[10] R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer, Y. Jararweh, and S. Maqsood, "Fake reviews detection: A survey," *IEEE Access*, vol. 9, pp. 65771–65802, 2021.

[11] Y. Wu, E. W. Ngai, P. Wu, and C. Wu, "Fake online reviews: Literature review, synthesis, and directions for future research," *Decision Support Systems*, vol. 132, p. 113280, 2020.

[12] S. Ansari and S. Gupta, "Review manipulation: Literature review, and future research agenda," *Pacific Asia Journal of the Association for Information Systems*, vol. 13, no. 1, p. 4, 2021.

[13] L. Breiman *et al.*, "Statistical modeling: The two cultures," *Statistical science*, vol. 16, no. 3, pp. 199–231, 2001.

[14] O. Müller, I. Junglas, J. v. Brocke, and S. Debortoli, "Utilizing big data analytics for information systems research: challenges, promises and guidelines," *European Journal of Information Systems*, vol. 25, no. 4, pp. 289–302, 2016.

[15] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, eds.), vol. 27, Curran Associates, Inc., 2014.

[17] D. Saxena and J. Cao, "Generative adversarial networks (gans): Challenges, solutions, and future directions," *CoRR*, vol. abs/2005.00065, 2020.

[18] W. Nie, N. Narodytska, and A. Patel, "RelGAN: Relational generative adversarial networks for text generation," in *International Conference on Learning Representations*, 2019.

[19] J. Guo, S. Lu, H. Cai, W. Zhang, Y. Yu, and J. Wang, "Long text generation via adversarial training with leaked information," *CoRR*, vol. abs/1709.08624, 2017.

[20] L. Yu, W. Zhang, J. Wang, and Y. Yu, "Seqgan: Sequence generative adversarial nets with policy gradient," *CoRR*, vol. abs/1609.05473, 2016.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[22] A. Holtzman, J. Buys, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," *CoRR*, vol. abs/1904.09751, 2019.

[23] M. Caccia, L. Caccia, W. Fedus, H. Larochelle, J. Pineau, and L. Charlin, "Language gans falling short," *CoRR*, vol. abs/1811.02549, 2018.

[24] G. Tevet, G. Habib, V. Shwartz, and J. Berant, "Evaluating text GANs as language models," *CoRR*, vol. abs/1810.12686, 2018.

[25] D. Zhang, L. Zhou, J. L. Kehoe, and I. Y. Kilic, "What online reviewer behaviors really matter? effects of verbal and nonverbal behaviors on detection of fake online reviews," *Journal of Management Information Systems*, vol. 33, no. 2, pp. 456–481, 2016.

[26] R. Barbado, O. Araque, and C. A. Iglesias, "A framework for fake review detection in online consumer electronics retailers," *Information Processing & Management*, vol. 56, no. 4, pp. 1234–1244, 2019.

[27] A. Dürr, M. Griebel, G. Welsch, and F. Thiesse, "Predicting fraudulent initial coin offerings using information extracted from whitepapers," in *Proceedings of the 28th European Conference on Information Systems (ECIS)*, pp. 1–16, 2020.

[28] E. Kauffmann, J. Peral, D. Gil, A. Ferrández, R. Sellers, and H. Mora, "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*, vol. 90, pp. 523–537, 2020.

[29] D. Plotkina, A. Munzel, and J. Pallud, "Illusions of truth—experimental insights into human and algorithmic detections of fake online reviews," *Journal of Business Research*, vol. 109, pp. 511–523, 2020.

[30] Y. K. Huang, W. I. Yang, T. M. Lin, and T. Y. Shih, "Judgment criteria for the authenticity of internet book reviews," *Library & Information Science Research*, vol. 34, no. 2, pp. 150–156, 2012.

[31] T. Fagni, F. Falchi, M. Gambini, A. Martella, and M. Tesconi, "Tweepfake: About detecting deepfake tweets," *PLOS ONE*, vol. 16, pp. 1–16, 05 2021.

[32] W. Zhong, D. Tang, Z. Xu, R. Wang, N. Duan, M. Zhou, J. Wang, and J. Yin, "Neural deepfake detection with factual structure of text," *CoRR*, vol. abs/2010.07475, 2020.

[33] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending against neural fake news," *CoRR*, vol. abs/1905.12616, 2019.

[34] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," *CoRR*, vol. abs/1705.07874, 2017.

[35] O. Loyola-González, "Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view," *IEEE Access*, vol. 7, pp. 154096–154113, 2019.

[36] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 0210–0215, 2018.

[37] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[38] W. Samek, T. Wiegand, and K. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *CoRR*, vol. abs/1708.08296, 2017.

[39] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[40] E. Winter, "Chapter 53: The shapley value," in *Handbook of Game Theory with Economic Applications*, vol. 3, pp. 2025–2054, Elsevier, 2002.

[41] J. Tritscher, M. Ring, D. Schlr, L. Hettinger, and A. Hotho, "Evaluation of post-hoc xai approaches through synthetic tabular data," in *Foundations of Intelligent Systems* (D. Helic, G. Leitner, M. Stettinger, A. Felfernig, and Z. W. Raś, eds.), (Cham), pp. 422–430, Springer International Publishing, 2020.

[42] K. Peffers, T. Tuunanen, and B. Niehaves, "Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research," *European Journal of Information Systems*, vol. 27, no. 2, pp. 129–139, 2018.

[43] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Information Fusion*, vol. 36, pp. 10–25, 2017.

[44] G. G. Chowdhury, "Natural language processing," *Annual Review of Information Science and Technology*, vol. 37, no. 1, pp. 51–89, 2003.

[45] L. Zhou, J. Burgoon, J. Nunamaker, and D. P. Twitchell, "Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications," *Group Decision and Negotiation*, vol. 13, pp. 81–106, 2004.

[46] M. Siering, J.-A. Koch, and A. V. Deokar, "Detecting fraudulent behavior on crowdfunding platforms: The role of linguistic and content-based cues in static and dynamic contexts," *Journal of Management Information Systems*, vol. 33, no. 2, pp. 421–455, 2016.

[47] S. Humpherys, K. Moffitt, M. Burns, J. Burgoon, and W. Felix, "Identification of fraudulent financial statements using linguistic credibility analysis," *Decision Support Systems*, vol. 50, pp. 585–594, Feb. 2011.

[48] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, p. 993–1022, Mar. 2003.

[49] C. Jacobi, W. van Atteveldt, and K. Welbers, "Quantitative analysis of large amounts of journalistic texts using topic modelling," *Digital Journalism*, vol. 4, no. 1, pp. 89–106, 2016.

[50] G. Shmueli and O. R. Koppius, "Predictive analytics in information systems research," *MIS quarterly*, pp. 553–572, 2011.

[51] G. Forman *et al.*, "An extensive empirical study of feature selection metrics for text classification.," *J. Mach. Learn. Res.*, vol. 3, no. Mar, pp. 1289–1305, 2003.

[52] N. A. Patel and R. Patel, "A survey on fake review detection using machine learning techniques," in *4th International Conference on Computing Communication and Automation (ICCCA)*, pp. 1–6, 2018.

[53] F. Abri, L. F. Gutiérrez, A. S. Namin, K. S. Jones, and D. R. W. Sears, "Fake reviews detection through analysis of linguistic features," *CoRR*, vol. abs/2010.04260, 2020.