Journal of Retailing and Consumer Services 64 (2022) 102771

Contents lists available at ScienceDirect

Journal of Retailing and Consumer Services

journal homepage: www.elsevier.com/locate/jretconser

Creating and detecting fake reviews of online products

Joni Salminen [a,b,*], Chandrashekhar Kandpal [c], Ahmed Mohamed Kamel [d], Soon-gyo Jung [a], Bernard J. Jansen [a]

[a] Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar
[b] Turku School of Economics at the University of Turku, Turku, Finland
[c] Jaypee Institute of Information Technology, Noida, India
[d] Cairo University, Cairo, Egypt

ARTICLE INFO

Keywords:
Fake reviews
Detection
e-commerce
eWOM
Marketing

ABSTRACT

Customers increasingly rely on reviews for product information. However, the usefulness of online reviews is impeded by fake reviews that give an untruthful picture of product quality. Therefore, detection of fake reviews is needed. Unfortunately, so far, automatic detection has only had partial success in this challenging task. In this research, we address the creation and detection of fake reviews. First, we experiment with two language models, ULMFiT and GPT-2, to generate fake product reviews based on an Amazon e-commerce dataset. Using the better model, GPT-2, we create a dataset for a classification task of fake review detection. We show that a machine classifier can accomplish this goal near-perfectly, whereas human raters exhibit significantly lower accuracy and agreement than the tested algorithms. The model was also effective on detected human generated fake reviews. The results imply that, while fake review detection is challenging for humans, "machines can fight machines" in the task of detecting fake reviews. Our findings have implications for consumer protection, defense of firms from unfair competition, and responsibility of review platforms.

1. Introduction

The "phenomenon of fake" is taking over marketing. Major drivers for this are (a) the rapid technological development that enables the creation of artificial consumer-facing outputs, such as deepfakes (Floridi, 2018; Jan et al., 2020; Tolosana et al., 2020), and (b) the marketplace evolving around these artificial outputs, related to fake creation, detection, and mitigation (Hajek and Henriques, 2017). Among the most impactful artificial marketing outputs are fake product reviews — also known as 'fake reviews,' 'deceptive reviews,' 'deceptive opinion spam,' 'review spam,' or 'review fraud' — that pass as real ones. To this end, studying fake reviews has been suggested as one of the primary agenda items in digital and social media marketing research (Dwivedi et al., 2020). Online product reviews, as a form of electronic Word-of-Mouth (eWOM), are major drivers in influencing consumers' purchase decisions (Duarte et al., 2018; Endo et al., 2012; Kaushik et al., 2018; Sandra MC Loureiro and Javier Miranda, 2018; Tran and Strutton, 2020). In the United States, more than 80% of consumers indicate they use online reviews before purchasing a product (Smith and Anderson, 2016). As reviews are among the most influential factors on consumers' buying behavior, fraudulent actors are tempted to hire writers who specialize in or use automated methods for generating fake reviews to enhance the attractiveness of their products and services, or to degrade competitors' reputation.

Fake reviews can be created in two main ways. First, in a (a) *human-generated way* by paying human content creators to write authentic-appearing but not real reviews of products — in this case, the review author never saw said products but still writes about them. Second, in a (b) *computer-generated way* by using text-generation algorithms to automate the fake review creation. Traditionally, human-generated fake reviews have been traded like commodities in a "market of fakes" (He et al., 2021) – one can simply order reviews online in a given quantity, and human writers would carry out the work. However, the technological progress in text generation – natural language processing (NLP) and machine learning (ML) to be more specific – has incentivized the automation of fake reviews, as with generative language models, fake reviews could be generated at scale and a fraction of the cost compared to human-generated fake reviews.

This issue is important for marketing and e-commerce domains for three main reasons. First, (a) *fake reviews may erode consumer trust in*

* Corresponding author. Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar.
E-mail address: joolsa@utu.fi (J. Salminen).

https://doi.org/10.1016/j.jretconser.2021.102771
Received 3 May 2021; Received in revised form 10 August 2021; Accepted 9 September 2021
Available online 20 September 2021

online reviews as a whole, which would signify a major market decline. Sincere consumers write reviews to share their experiences, either positive or negative. Hence, truthful reviewing renders a valuable service in the marketplace (Munzel and Kunz, 2014), as the information in these reviews provides a signal of quality for other consumers. A truthful marketplace for reviews is also in the interest of companies, as they can receive authentic feedback from customers that can be analyzed to improve products and services. If fake reviews were to permeate the marketplace at scale, this would risk systematically degrading source credibility (Ismagilova et al., 2020) of online reviews in general. The consequence might be adverse selection, a process in which consumers are unable to distinguish good reviews from bad ones (Akerlof, 1970).

Second, (b) *fake reviews can influence a product's ranking either positively (when the fake review is positive) or negatively (when the fake review is negative)*. This is because online marketplaces' algorithms use reviews as a signal to determine a product's ranking among other products in the same category (Gobi and Rathinavelu, 2019). Therefore, fake reviews can result in unfair competition, where a product's ranking is artificially inflated or deflated (He et al., 2021). This means that fake reviews can be weaponized – an unethical firm may generate an influx of negative reviews about its rival. Flooding the market with such reviews can cause the ranking algorithms of online platforms (e.g., Google, Facebook, Twitter, Yelp, TripAdvisor, Amazon) to lower the visibility of the attacked firm. It is essential to detect and prevent such effects from taking place in order to protect firms from unfair competition.

Third, (c) *the impact of fake reviews is not only cosmetic or reputational but involves a financial cost as well*. For example, Luca (2011) estimates that a one-star decrease in a company's Yelp rating results in a five to nine percent decrease in revenue.

Because of these three main reasons, fake reviews risk imposing severe negative impacts on firms' profits and consumers' well-being, becoming a "dangerous prospect for online users" (p. 1) (Ahmed et al., 2018) and potentially having a major impact on the online marketplace due to the general prominence of reviews (Crawford et al., 2015). Through platforms' algorithmic decision-making mechanisms, reviews have become an integral part of social media and e-commerce shopping experiences. Therefore, the quality of reviews (a manifestation of eWOM) is essential to brands, e-commerce sites, social media platforms, and other stakeholders with a vested interest in online business. Consequently, identifying fake reviews is an issue of the utmost importance. In this research, we address fake review detection through the following research questions (RQs):

- **RQ1:** How realistic (i.e., high-quality) are reviews that current text generation algorithms produce? (i.e., can text generation fool humans?)
- **RQ2: (a)** Can a machine detect a fake review generated by another machine? **(b)** Does a machine classifier do so better than a human?)
- **RQ3:** Can a machine detect a fake review generated by humans?

Because machine generation of fake reviews is becoming more common (Floridi, 2018; Jan et al., 2020; Tolosana et al., 2020), it is important to have a clear understanding of the capability of current technology to generate such reviews. It is also important to detect fake reviews to preserve the credibility of the marketplace (Ismagilova et al., 2020). Therefore, knowledge on fake review detection by machines or humans informs academics and firms about the ability of current text generators' ability to create convincing fake reviews and of the ability of machine and human classifiers to detect this form of deception in an effective way.

2. Related literature

2.1. Conceptual underpinnings

A *fake review* is a review written or generated without any actual experience of the product or service being reviewed (Lee et al., 2016). Because it is "written" or "generated," a fake review can be created manually by a human writer or automatically by a computer program. As such, technological progress is a major enabler of fake reviews, as it creates both opportunities and incentives for manipulating consumer decisions (Ahmed et al., 2018). On the one hand, advances in natural language generation (Floridi and Chiriatti, 2020) provide opportunities for large-scale *production* of fake reviews – technology may be reaching the pinnacle in which it becomes virtually impossible for human readers to detect if a given piece of text is written by a real person. On the other hand, online platforms provide a *distribution* channel for large-scale diffusion of fake reviews. As literally millions of consumers read online reviews at any given moment, there is an incentive to exploit this route of persuasion at scale. Therefore, fake reviews potentially benefit from economies of scale and scope, which accentuates the challenge of developing contrary measures for this type of misinformation.

Opinion spamming is a concept similar to fake review, defined as writing false opinions (i.e., opinion spam) to influence other online users (Ahmed et al., 2018), or as "fictitious opinions that have been deliberately written to sound authentic" (Ott et al., 2011) (p. 1). Ott et al. (2011) contrast opinion spam to imaginative writing, postulating that research from computational linguistics and psychology is beneficial for detecting opinion spam (Ott et al., 2011). Another concept similar to fake reviews is *incentivized review*, which refers to reviews obtained via a marketing campaign – e.g., by obtaining influencer endorsements (Petrescu et al., 2018) or by offering consumers a particular product in exchange for a review (Costa et al., 2019). However, these endorsements differ from fake reviews in a central way: fake reviews are typically written by anonymous users or those using pseudonyms, whereas incentivized reviews are typically traced back to an influencer or a "real person". This means that the plausibility of incentivized reviews may be higher than that of fake reviews, as the incentivized authors write the reviews under their own names. For an influencer, reputation is important, and they would typically avoid writing misleading reviews in order to maintain the audience's trust (Kaabachi et al., 2017). Nonetheless, the similarity is that both fake and incentivized reviews are written partially or completely based on financial incentives rather than genuine (dis)liking of the product.

One aspect that is important to underline is that the desired impact of fake reviews can be to either enhance or damage an organization's reputation, and the source of the fake reviews can be either the organization itself or third parties, such as competitors or vengeful customers. Acknowledging the variability in motives is important for understanding the full scope of the phenomenon – fake reviews not only constitute praises for a given product but can also include attacks against rival brands. A positive review of a target brand, company, or product is designed to attract more customers and increase sales, whereas a negative review aims to tarnish the target's reputation and decrease sales (Shivagangadhar et al., 2015).

2.2. Fake review detection

A basic approach to fake review detection is to analyze reviews manually. This approach is based on the premise that humans can detect when other humans behave in fraudulent ways — i.e., knowledge of the "psychology of lie" (DePaulo et al., 1996). The advantage of careful perusing of fake reviews is that it affords developing heuristic rules that can be understood and interpreted. For example, Costa et al. (2019) identified a set of rules to distinguish between incentivized and non-incentivized reviews, including the review's length, sentiment, and helpfulness rate. Jindal and Liu (2008) identified general patterns in online reviews, such as only a small number of reviews per user and product, and that the reviews rarely garner much feedback. If a review differs from these patterns, it may have a higher probability of being fake. Filieri (2016) investigated how humans assess the trustworthiness of an online review and found factors, such as the review's content and

writing style, including the presence of pictures, length, degree of detail, and extreme positivity or negativity.

One eminent challenge with the use of heuristic rules is that they may not always be accurate – for example, Sandulescu and Ester (Sandulescu and Ester, 2015) raise the point of singleton spammers, i.e., fake reviews written by users with a not exceptionally high review count. If using the number of written reviews as a signal, singleton spammers may be left undetected. Another challenge is that once the rules of fake detectors become common knowledge, spammers adapt to them and change their behavior, rendering the rules invalid – this is a variant of Goodhart's Law ("When a measure becomes a target, it ceases to be a good measure.") (Mattson et al., 2021). These challenges can explain the general deficiency of human performance for predicting fake reviews. For example, Ott et al. (2011) recruited humans to judge if a review was fake or not, finding that the highest accuracy for a human (65%) was substantially lower than for an ML model (86%). Plotkina et al. (2020) found humans had a 57% detection accuracy, even when conditioned by information cues about fake reviews. In another study by Sun et al. (2013), human accuracy was 52%, suggesting that it is very difficult for humans to separate fake reviews from real ones. Apparently, the heuristics applied by people are not effective against a range of deceptive tactics deployed in fake reviews.

Another challenge of manual detection is that the number of online reviews is growing at an exponential rate. For example, TripAdvisor reported having more than 200 million reviews (Crawford et al., 2015). Given that a single product may receive thousands of reviews (Algur et al., 2010), and reviews exist for millions of objects (i.e., companies, products, service providers), manual methods simply do not scale for assessing such a volume of reviews. Hence, researchers see the use of automatic methods as a potential solution to the fake review detection problem. Using automation, an algorithm mines the data samples for patterns, some of which may not even be interpretable for humans. Cardoso et al. (2018) divide automated methods into (a) *content-based detection* (focused on textual content in the reviews), (b) *behavior-based detection* (focused on atypical and suspicious behaviors), (c) *information-based detection* (focused on product characteristics), and (d) *spammer group detection* (focused on identifying connections between reviewers).

Automatic detection that applies NLP techniques predominantly focuses on reviews as textual data, hence emphasizing lexical features (i.e., attributes derived from text), such as keywords or -phrases, n-grams, punctuation, semantic similarity, latent topics, and indicators of linguistic styles (Jindal and Liu, 2008; Mihalcea and Strapparava, 2009; Mukherjee et al., 2013; Ott et al., 2011; Sandulescu and Ester, 2015). Another stream of literature focuses on non-textual predictive features, such as user IDs, user location, number of reviews generated by a user, and other potentially suspicious behaviors (Mukherjee et al., 2013). As typical for classification tasks, approaches that combine different types of features tend to perform better at fake review detection (Ott et al., 2011). The key takeaway is that features may include both textual and non-textual information (a thorough review of different features is provided in (Crawford et al., 2015)). Our study is positioned in the lexical analysis of fake reviews, as we experiment with language models and text classifiers. This line of work can be considered as an alternative to heuristic- and behavior-based fake review detection.

Combinations of the two approaches (i.e., manual and automatic), while theoretically possible, are rare (Munzel, 2016). Among those rare examples, Munzel's study (Munzel, 2016) emphasizes the role of sharing not only textual but also contextual information (e.g., identity disclosure, consensus of reviews) with human detectors to assist them in detecting fake reviews. Harris (2019) devised a hybrid approach in which human evaluators were given information about psycho-linguistic features that were extracted algorithmically, along with the decisions by two ML classifiers. The humans had the option to agree or disagree with the machine decision – using this hybrid approach, they improved machine performance by 0.2 percentage points, showing that human participation can result in a marginal (but statistically significant) improvement over a purely machine-based approach.

The takeaway is that fake review detection varies from completely automated to completely manual procedures. It is important to note that even if the decision of whether a given review is fake or not is made automatically by a classification model, a human or a group of humans has always had a role in training that model — through dataset creation, data pre-processing, feature engineering, and hyperparameter selection. Therefore, what is termed as "automatic" fake review detection is still characterized by activities of human labor, which may involve some degree of biased thinking (Kirkpatrick, 2016). In practice, this implies that the human choices during the classifier development should be reported in a transparent way, possibly by making computational notebooks publicly available (as we do in this study; see "Limitations and Future Development").

2.3. Datasets

There are a handful of datasets for fake review detection, but each of these comes with its own difficulties. Jindal and Liu (2008) analyzed 10M reviews from Amazon.com with the goal of detecting various types of opinion spam. Even though they obtained reasonable performance – accuracy of 63% with text features and 78% with all available features – the validity of the results is hindered by the fact that they labeled completely duplicated and near-duplicated reviews automatically as fake reviews, even though such occurrences might arise from legitimate reasons as well. In other words, not all abnormal activity is necessarily fraudulent. In contrast to the study by Jindal and Liu (2008), we are completely certain that the fake reviews in our dataset are fake, as they did not exist before we generated them.

Ott et al. (2011) developed a dataset with 800 fake reviews and 800 truthful reviews. The truthful reviews were collected from TripAdvisor, representing the 20 most popular hotels in an American city. The fake reviews were written by 400 crowd workers recruited via Amazon Mechanical Turk. The workers were asked to make either a positive or a negative review for a given hotel that they had no experience with. The given hotels included the same 20 hotels as in the truthful reviews. Despite representing a considerable step forward in understanding fake reviews, the dataset by Ott et al. (2011) has two major limitations. First, the dataset size ($n = 1600$ reviews in total) is small for training effective text classifiers. Second, the researchers omitted reviews with less than 150 characters and those with less than five stars (maximum score) when collecting the dataset. However, as mentioned previously, fake reviews are not necessarily positive ones, and so the exclusion of reviews with less than five stars may not be appropriate. Similarly, the length distribution of fake reviews may extend to under 150 characters. We consider these aspects in our study as we generate a considerably larger dataset for fake reviews and ensure a proportional representation of reviews representing different ratings and lengths.

Yoo and Gretzel (2009) collected 42 fake and 40 truthful hotel reviews and compared psycho-linguistic differences among those reviews. Again, the dataset is too small for training effective ML classifiers to detect fake reviews at scale. Sandulescu and Ester (Sandulescu and Ester, 2015) obtained a dataset containing 9000 reviews labeled as fake or real reviews. The dataset was shared by an online company called Trustpilot, and the dataset includes four- and five-star reviews from 130 companies, limited to one-time reviewers only. Yet, the dataset has not been made publicly available, which hinders replication and further development. Moreover, this dataset is biased to positive reviews at the expense of detecting negative fake reviews. To this effect, we develop our own dataset of fake reviews that are publicly available and, as mentioned, use reviews from all rating levels when generating the dataset.

2.4. Summary and research gap

As fake reviews pose a pervasive and damaging problem, helping consumers and businesses differentiate truthful reviews from fake ones remains a vital but challenging task (Crawford et al., 2015). Fake review detection can combine manual efforts, supervised ML, and heuristic methods (Fontanarava et al., 2017). Some approaches in the literature focus solely on features extracted from the review text. Linguistic characteristics range from counting the frequency of words or n-grams (Viviani and Pasi, 2017) to more advanced approaches relying on distributional semantics (Lee et al., 2016). However, despite the progress made in detection studies, considerable challenges lie ahead. Classification performance needs improvement to keep up with text-generation algorithms. Datasets may not be appropriately devised, contain mislabeled instances, or are not made publicly available. The key takeaway from previous studies is that automatic fake review detection has been only partially successful. While one study cannot tackle all gaps, our study leverages state-of-the-art NLP technologies to generate a robust dataset for fake review detection and then compare manual (crowdsourcing) and automated (ML algorithm) performance to detect computer-generated fake reviews. We make our experiments available for future development.

3. Methodology

The procedure follows six main steps: (1) *Generate* sample reviews using two language models. (2) *Evaluate* the generated sample reviews using quantitative metrics and qualitative assessment. (3) *Choose* the best language model for the creation of a fake review dataset, which also contains the original, human-written reviews. (4) *Train* classifier algorithms to detect artificially generated reviews from the real ones. (5) *Recruit* crowd workers to annotate a sample of the original and fake reviews. (6) *Compare* the accuracy of crowd workers and the classification algorithms via statistical testing. Our approach is based on synthetic review generation, which constructs fake reviews based on pre-existing real reviews (Crawford et al., 2015). This approach is an alternative to identifying fake reviews from real ones and labeling them accordingly (Jindal and Liu, 2008) or paying human writers to create fake reviews (Ott et al., 2011). Applying this approach is important, as fake review creators are all the time searching for more efficient ways to generate fake reviews at scale, with minimum human involvement (Sun et al., 2013). To defend against these approaches, researchers need to experiment with computer-generated fake reviews and build classifiers based on synthetic outputs in order to mimic the behaviors of fake review attackers in the wild.

4. Fake review generation

4.1. Language models

We use two language models to generate fake product reviews. A language model learns to predict the probability of a sequence of words. More specifically, generative language models ("generators") can be optimized for two goals: (a) *open-ended generation*, where the generator has the "artistic freedom" to generate text that matches the language tendencies learned during the training and fine-tuning stages of model development; and (b) *purposeful generation*, where the model is expected to generate a specific piece of text that is strictly derived from the input. Translation is an example of purposeful generation – a sentence in Finnish should be generated in Korean while preserving the same meaning. The fake review generation task deals with open-ended generation.

4.1.1. ULMFit

Howard and Ruder (2018) introduced Universal Language Model Fine-tuning (ULMFiT) to facilitate transfer learning across NLP tasks. ULMFiT includes three main procedures: (1) language model pre-training, (2) model fine-tuning (i.e., discriminative fine-tuning), and (3) classifier fine-tuning. Because the model has already captured the general properties of the language during pre-training, it is proposed that a relatively minor tweaking (i.e., fine-tuning) is sufficient for adapting the model to a specific task – e.g., text classification or generation. ULMFiT represents a strong contribution in NLP, paving the way for more advanced transfer learning models, such as those based on transformers (Wolf et al., 2019) – including GPT-2. Therefore, we consider ULMFiT as a solid baseline for our experiments.

4.1.2. GPT-2

OpenAI's GPT-2 model was proposed by Radford et al. (2019). GPT-2 is a causal (unidirectional) transformer pre-trained using language modeling on a very large corpus of ~40 GB of text data from 8 million web pages. GPT-2 has 1.5 billion parameters, making it one of the most advanced models in the modern NLP. GPT-2 is trained with the objective of predicting the next likely word based on the previous words in the context. As a consequence of the large parameter space and diverse training data, GPT-2 is shown to generalize relatively well to various tasks in different domains (Budzianowski and Vulić, 2019; Floridi and Chiriatti, 2020), which leads us to believe that the model could be applicable for product review generation as well.

The choice of these two models was driven by two main factors:

Both models are based on transfer learning, which aims to mitigate the need for task-specific adjustments and train the model every time from the start. Given a source context in which the model is trained, the same model is used to achieve solid performance in other application contexts.

Both models are available as open-source code, which makes fine-tuning possible. Note that at the time of the study, GPT-3, an advanced version of GPT-2, was published; however, the weights of this model had not been made publicly available, so we could not apply that model in this study. (Regardless, GPT-2 performed exceedingly well, as shown by the results.)

The main difference between the models is that ULMFiT has long short-term memory (LSTM) as a backbone, while GPT-2 is a transformer-based model. In other words, the models represent different types of NLP architectures. Apart from this, both models rely on transfer learning, which is a major advancement in NLP (Devlin et al., 2018; Raffel et al., 2019). In transfer learning, a language model is first pre-trained on a large text corpus to learn the general aspects and properties of language. The pre-training needs to be performed only once, as the obtained pre-trained model can be reused as a basis for various downstream NLP tasks, including text generation and classification. With extensive pre-training, the model learns to understand the general properties of language and thus needs to only be fine-tuned to adapt to a specific context. Pretraining is particularly beneficial for small- and medium-sized datasets, including those used for research purposes.

4.2. Dataset

Before addressing the fake review generation, we must first collect a suitable dataset of product reviews that can be used for model fine-tuning. Data forms a vital component of any ML model, as it directly impacts the quality of results. In this study, we use the publicly available Amazon Review Data (2018) dataset, which is extensive and reputable (Ni et al., 2019). As suggested by the creators of the dataset, we use the k-core subsets for experimentation purposes. The data has been reduced to extract the k-core (i.e., a dense subset), such that each of the remaining users and items has k reviews each (see Fig. 1).

We leverage the Amazon dataset for two purposes. First, we train generative models on samples of multiple categories to compare the performance between ULMFiT and GPT-2 (i.e., model comparison).

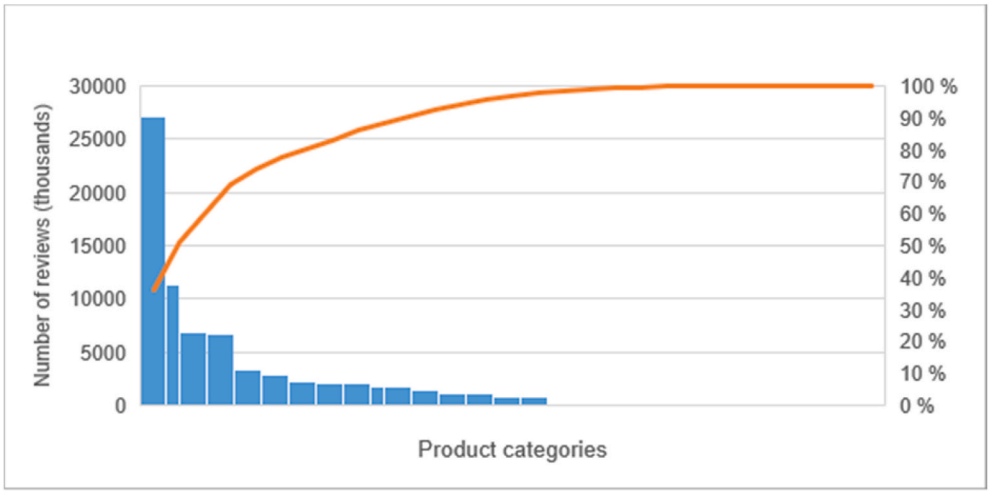

Fig. 1. A Pareto chart showing the distribution of product reviews by product category in descending order of frequency, with a cumulative line on a secondary axis as a percentage of the total. As can be seen, the dataset is highly imbalanced. However, for our experiments, this does not matter for two reasons: (a) even the smallest class ("Appliances") has 2277 samples, which is a good number, and (b) we apply stratified sampling in the generation to evenly generate reviews from each selected product category.

Second, we use samples from the Top-10 categories (based on review count) and train models on these individual categories to build a dataset of fake reviews for classification. For both purposes, we use an adequate number of samples to get a reliable picture of performance. For the model comparison, we use the following categories (the number of samples in parentheses): Beauty (n = 5269); Automotive (n = 1,711,519); Gift cards (n = 2972); Magazine subscriptions (n = 2375); Fashion (n = 3176); CDs and vinyl (n = 1,443,755); Grocery and gourmet food (n = 1,143,860); Musical instruments (n = 231,392); Appliances (n = 2277); Cell phones and accessories (n = 1,128,437); Industrial and scientific (n = 77,071); Office products (n = 800,357); Arts, crafts, and sewing (n = 494,485); Digital music (n = 169,781); Luxury and beauty (n = 34,278); and Patio, lawn, and garden (n = 798,415). In total, 40,000 samples (~2500 per category) from these categories were extracted for model training, and 10,000 (625 per category) were extracted for model validation.

4.3. Review-generation experiments

Our experiments showed that GPT-2 outperformed ULMFiT on all relevant metrics. Therefore, we use GPT-2 to build the fake review dataset for the classification task. As can be seen from Table 1, GPT-2 outperforms ULMFiT. *Training loss* indicates the predictive error training data (i.e., data the model has not seen). *Validation loss* indicates the predictive error on validation data (i.e., the dataset used for hyperparameter optimization). *Perplexity* is used to judge how good a language model is (Clarkson and Robinson, 1997). In simple terms, a perplexity of 100 means that whenever the model is predicting the next word, it is as confused as if it had to pick between 100 different words. *Training time* indicates the duration it takes to train the model (hh:mm:ss). In general, if two models have equal performance, the one that can be trained faster is considered better. Therefore, for all these metrics, a lower value implies better performance.

Based on a manual review (see samples in Table 2), we observed that both models tend to generate partial sentences (e.g., the first two examples for both GPT-2 and ULMFiT in Table 2). To mitigate this, a post-processing approach was devised. First, if the generated review ends with a full stop (i.e., "."), we accept the review as it is. If the last sentence of the generated review has either one, two, or three words, we simply omit those words and accept the rest of the review as the generated review. If the last sentence of the generated review has more than three words, the sentence typically starts to make sense, so we accept the review as it is. This can be understood with an example. Let us say the

Table 1
Performance comparison metrics (lower is better). GPT-2 outperforms ULMFiT on all relevant metrics after only a training epoch – even after 10 epochs, ULMTFiT is unable to catch up. All the experiments were performed on an Intel® Core i7-8750H CPU @ 2.20 GHz machine running Ubuntu 20.04. We used the NVidia GeForce 1060 GPU for the majority of the model training and fine-tuning.

epoch	training loss	validation loss	perplexity	training time
GPT-2				
0	3.773	3.673	39.355	1:28:52
ULMFiT				
0	4.119	3.975	53.260	1:49:43
1	4.131	4.017	55.547	1:53:25
2	4.092	4.005	54.885	1:56:59
3	4.072	3.982	53.613	1:52:29
4	4.035	3.947	51.774	1:47:57
5	3.976	3.911	49.953	1:48:03
6	3.942	3.873	48.099	1:48:13
7	3.857	3.836	46.351	1:48:01
8	3.821	3.813	45.272	1:47:59
9	3.815	3.808	45.038	1:47:55

Table 2
Example reviews from three product categories generated by GPT-2 and ULMFiT. Based on a manual review, GPT-2 seems to produce more believable reviews.

	GPT-2	ULMFiT
Toys and Games	Amazing toy material is soft and sturdy. I bought this for my 4 year old son as a gift for Christmas. He loves it.My daughter got this for her 4th birthday and loves it. It's sturdy.Very well made	Amazing toy material is an excellent choice for our 6 year old granddaughter. She loves it ! This is the best Food set we've bought for her. She uses all the animals and matching dough and we have bought her some water food and other food and
Tools and Home improvement	Amazing tool. I use it every day just to see what the weather is doing and I have no complaints. I really like the price and it works well. I would recommend it!This is a very good kit. It would	Amazing tool ! Bright enough to see what you need and features, especially for the graphic of the yard ! Bulky but still pleased with the tool. Works great for screwdrivers and scissors. We purchased a different style for our new house in
Pet supplies	Taste is good but packaging was poor but my dogs love it. My cats love the taste. They like the taste and stay away from the other cat food. I'm also noticed that they prefer the other food that I've tried.	Taste is good but packaging was poor but my cats wouldn't eat it. Anyway, my picky eater did not like this food. Seemed to be the same as the other flavors but she loved it too. These are perfect for the purpose of removing the litter box from confusion.

generated review was, "He sort of likes it, but it's not very comfortable for him. My dog is". So, after post-processing, this generated review will look like, "He sort of likes it, but it's not very comfortable for him." Apart from this simple post-processing step, we include the reviews in the dataset in the exact format that the GPT-2 generates them.

4.4. Review generation and creation of the fake reviews dataset

To generate the reviews, we adopt a stratified sampling approach, meaning that we randomly select an equal proportion of reviews to generate per category. For this part, due to mitigating computational complexity, we choose to sample the Top-10 Amazon categories with the most product reviews (i.e., Books; Clothing, Shoes, and Jewelry; Home and Kitchen; Electronics; Movies and TV; Sports and Outdoors; Kindle Store; Pet Supplies; Tools and Home Improvement; Toys and Games). Reviews from these categories account for 88.4% of the reviews in the baseline dataset, thereby representing the baseline dataset (Amazon product reviews) reasonably well.

For each product category, we generate 2000 reviews using a fine-tuned GPT-2 language model. Technical details of the fine-tuning are omitted from this report but can be reviewed and accessed in the computational notebook provided as supplementary material.[1] In brief, GPT-2 takes w initial words from each sampled review and generates a remainder of the review automatically. Here, we set $w = 5$, as this parameter value falls in the typical range in text generation tasks (Liang and Zhu, 2018). Based on reviewing the reviews' length distribution (i.e., how many words the reviews contained in the whole dataset), we create discrete buckets with one-word intervals in the range of 10…350 words that we adopt as the target sentence length for the generated reviews (see Fig. 2a). We manage the proportions as per the original distribution in the sample; for example, if the proportion of 50-word reviews in the Amazon dataset is 0.5% of the total reviews, then 0.5% of the generated samples will also be 50 words of length.

In other words, the generated reviews will now satisfy two conditions: (a) *they are approximately equally divided across the most popular product categories*, and (b) *they account for different lengths of reviews*. As shown in Fig. 2b, this sampling strategy will also result in including a similar number of reviews in different rating levels – i.e., those receiving low ratings (one star) to those receiving high ratings (five stars) – as the sampled length range includes reviews from each rating level. Fake review detection across different rating levels is important to account for both positive and negative review types. Positive reviews are praising reviews generated with the intent of boosting a product's standing in rankings. In contrast, negative reviews are generated with the malicious intent of decreasing a competitor's reputation in the marketplace.

The resulting dataset includes 20,000 artificially-generated (fake) reviews. It also includes 20,000 (real) reviews written by humans (i.e., original samples from the Amazon dataset). Hence, there are 40,000 reviews in total. In general, this is a good number of samples for a text classification task, as binary text classification has been completed with a considerably smaller sample size (Salminen et al., 2018). With our approach, the real reviews are randomly picked, and the fake reviews are randomly generated based on the original dataset and the stratified sampling approach. As a result, the dataset can be used for the fake review detection task, where the two classes are *computer-generated reviews* (CG) and *original reviews* (OR). In the next step, we will train ML classifiers to distinguish CGs from ORs. Technically, as there are two classes, this will be a binary classification task.

[1] https://github.com/joolsa/FakeReviews/blob/main/nbs/generation/1_GPT2.ipynb

5. Fake review detection

5.1. Dataset description

Fig. 3 shows information about the fake reviews dataset. The dataset can be used toward the development of ML algorithms for fake-review detection. We experiment with various ML classifiers to address our RQs and to provide performance benchmarks.

5.2. Algorithm selection

A baseline algorithm or model in NLP tasks is typically chosen to represent the standard performance. In our case, we include two types of baseline models that we aim to outperform with our own fine-tuned model: (i) the support *vector machine* (SVM) algorithm, and (ii) the *OpenAI fake detection model*. SVM is a classic baseline algorithm used in NLP tasks (including fake review detection (Harris, 2019)) due to its robust performance (François and Miltsakaki, 2012). It applies kernel equations for linear classification on non-linear data, constructing a hyperplane on a multi-dimensional space that enables dividing the data into two classes (fake and real reviews, in this case). The specific variation of SVM that we use is called NBSVM (i.e., SVM with Naïve Bayes features) (Wang and Manning, 2012). The features used by NBSVM can be interpreted by humans (i.e., they can be traced back to individual words, symbols, and combinations thereof), a property that we will exploit at later stages.

The OpenAI model, in turn, is specifically developed toward the detection of fake reviews[2] and, as such, is a logical choice for a baseline approach. The model is based on the idea of fine-tuning a RoBERTa (Robustly Optimized BERT Pretraining Approach) model for the specific classification task. The RoBERTa model, proposed by Liu et al., in 2019 (Liu et al., 2019), builds upon Google's BERT (Bidirectional Encoder Representations from Transformers) model released in 2018 (Devlin et al., 2018). The model modifies key hyperparameters of BERT, removing the next-sentence pre-training objective and training with much larger mini-batches and learning rates. Through this, it provides performance gains relative to standard NLP classifiers (Liu et al., 2019). Overall, RoBERTa features are numerical vectors that cannot be interpreted by humans (but that tend to outperform simpler features).

Our model is inspired by OpenAI's idea of fine-tuning the RoBERTa model for a specific purpose. Therefore, we fine-tune RoBERTa on our dataset consisting of real reviews (OR) from the Amazon dataset against our generated reviews (CG) from the fine-tuned GPT-2 in the previous research phase. The intuition is that RoBERTa is a masked and non-generative language model that does not share the same architecture or the same tokenizer as GPT-2. This becomes important because, during the generation phase, we generated our reviews by fine-tuning GPT-2. As such, using a different architecture for the classifier intuitively makes it more independent from the GPT-2 style of text generation. We call this fine-tuned model *fakeRoBERTa*. The technical details of the fine-tuning, including parameter selection, are omitted due to their technical nature but can be seen and accessed in the shared computational notebooks for replication.[3]

5.3. Experimental procedure

The three models were trained and evaluated using the typically applied 80/20 split (Bell, 2014), which means that 80% of the dataset (n = 32,000) is used for training the model, and the remaining 20% (n = 8000) is held out for evaluation purposes. In other words, the test set

[2] https://github.com/openai/gpt-2-output-dataset/blob/master/detection.md

[3] https://github.com/joolsa/FakeReviews/blob/main/nbs/classification/2_roberta_finetune_amazon_reviews.ipynb

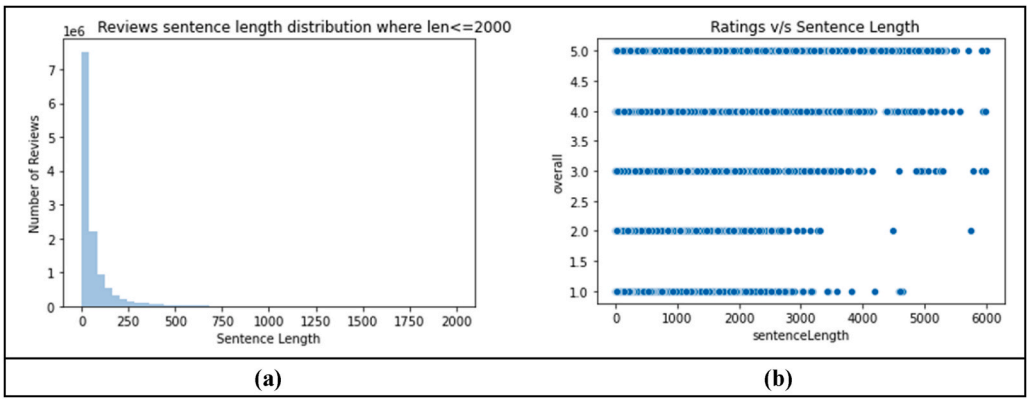

Fig. 2. **(a)** Sentence length distribution (M = 305.6 characters, SD = 307.0). The distribution is skewed, and the number of reviews drops considerably after 350 words. In fact, length of 2000 or more words is only ~2.5% of the overall reviews. **(b)** Review rating does not correlate with sentence length, as can be seen from scatterplot; every rating level (1…5) has reviews of all lengths. As such, there is no need to stratify the sample by rating.

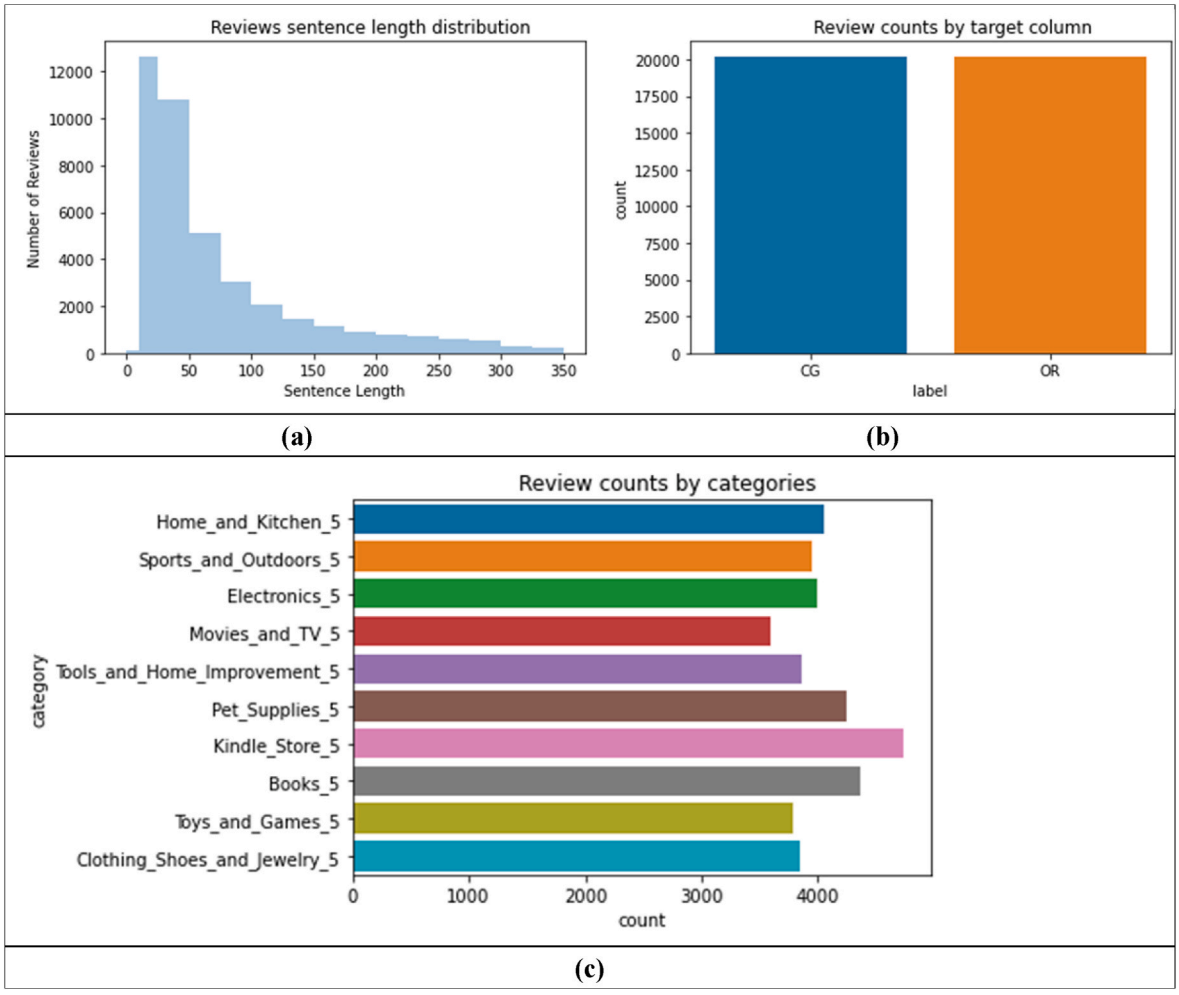

Fig. 3. Properties of the fake reviews dataset: **(a)** The length of the reviews follows the distribution in the original dataset. **(b)** The two classes are equally balanced. Since the number of samples per class is balanced, the classification is fair without the need to consider class imbalance in algorithm choice or data adjustment (i.e., under- or over-sampling or data augmentation is not needed). **(c)** There is a roughly equal number of reviews from each Top-10 product category. Hence, the classifier has a good chance of learning class separation across different product categories, thus increasing its generalizability.

contains samples the model was not exposed to during the model training. We evaluate model performance using several common ML metrics, including *precision*, which measures the proportion of cases identified as positive that were actually correct; *recall*, which measures the proportion of actual positives that were identified correctly; and *F1-score*, which is the harmonic mean of precision and recall and measures the model's overall ability to detect true cases (i.e., fake reviews in our case).

6. Results

The results in Table 3 show that our model has a precision of 0.97. In other words, when the model predicts a review is fake, it is correct 97% of the time. Similarly, our model's recall is 0.97 — it correctly identifies 97% of all fake reviews. Comparing the F1-scores, we see that fakeRoBERTa achieves an 18.3% performance increase over the OpenAI model and a 2.1% performance increase over the NBSVM model. Therefore, this new model outperforms both baselines. The performance gains are further illustrated in Table 4, showing the results per product category.

Fig. 4 shows the stability of classifiers' predictions by review length. The performance is more inconsistent with shorter reviews. This is most likely because short reviews contain less information for the model to judge – consider a review such as "*Ok, it works great!*" From a classification standpoint, it is difficult, if not impossible, to tell whether the review is written by a human or a machine. Consider, in turn, a review like this: "*Ok, it works great! this is so great.i wish i had it every day*." In this example, the use of ".i" (without space) is one indicator that is more common for machine-generated text – i.e., a type of grammatical error. The more text the classifier sees, the more likely it is that such cues become available for its final decision. The classifier makes the final decision based on a large number of cues: words, expressions, and their combinations. So, naturally, the more cues there are, the more likely the decision is correct.

Finally, Table 5 shows a variable importance analysis on the NBSVM model's features, which are human interpretable. The results imply that the model is able to implicitly detect some grammatical rules, such as a space following full stop and "i" (i.e., ". i …"). In contrast, fake reviews would be more likely to contain a pattern that ignores space in this context (i.e., ".i …"). However, this rule is not immutably understood by the model, as evidenced by the fact that the similar expression of ". it" is an indicator of a computer-generated review. This discrepancy can be explained so that the generator likely becomes biased in its fine-tuning process, over-learning expressions such as ".i …" and ". it" and then regularly using them but not their variations. The classifier model, in turn, is able to detect this bias and partially utilize it as information to distinguish fake reviews from real ones.

Note that this does not mean that the reviews written by people with poor literary skills would always be flagged as fake by the model. The classifier's decision is a question of accumulating evidence. More specifically, ".i" is not the only indicator for the decision – there are hundreds or possibly thousands of similar indicators, either supporting the decision of the review being fake or being against such a decision. Intuitively, the value of each indicator is summed up to make the final decision. Each indicator will have a (proverbial) weight. For example, the weight of ".i" can be quite high for fake reviews, but if that is the only indicator that supports the review being fake, then it is still very possible that the classifier will classify the sample as a real review. In other words, accumulating evidence helps the classifier to make a correct decision on whether to classify the review as fake or not.

Table 3
Predictive performance of the classification models. Highest performance highlighted.

Model	Precision[a]	Recall[b]	F1-score[c]
OpenAI	0.83	0.82	0.82
NBSVM	0.95	0.95	0.95
fakeRoBERTa	0.97	0.97	0.97

[a] Precision = True positives/(True positives + False positives).
[b] Recall = True positives/(True positives + False negatives).
[c] F1 = (2 * precision * recall)/(precision + recall).

Table 4
Predictive accuracy by product category with highest performance highlighted. Apart from one category (Home and Kitchen), our fine-tuned RoBERTa model achieves the best performance.

Category	OpenAI	NBSVM	fakeRoBERTa
Books	84.282	94.077	97.039
Clothing, Shoes, and Jewelry	78.913	94.955	95.990
Electronics	82.772	95.596	98.446
Home and Kitchen	82.071	96.465	96.086
Kindle Store	86.383	94.444	96.296
Movies and TV	83.356	93.398	96.561
Pet Supplies	81.517	93.720	96.564
Sports and Outdoors	80.048	95.913	97.476
Tools and Home Improvement	80.176	95.107	96.236
Toys and Games	79.973	93.767	94.960

6.1. Model generalizability

To investigate the generalizability of our fakeRoBERTa classifier, we applied the classifier to an independent dataset. This dataset was the one published by Ott et al. (2011), referred to as the "Deceptive Reviews Dataset". As explained in Section 2.3, the dataset has 1600 reviews, of which 800 (50%) are labeled as deceptive and 800 (50%) as truthful. To match these classes with our study, we consider the deceptive class as fake reviews, and the truthful class as original reviews.

Ott's dataset is particularly useful as its reviews are human-generated, which affords us the possibility to test our classifier, trained with machine-generated reviews, on a human-generated dataset. There may be linguistic differences between human- and computer-generated fake reviews that could make the model specifically applicable for the latter but not the former. The ideal fake review detector is able to identify both human- and computer-generated deceptive reviews (Sun et al., 2013).

We compare the classifiers using the AUC-ROC curve. A receiver operating characteristic (ROC) curve visualizes a binary classifier's predictive ability when varying its discrimination threshold (Bradley, 1997). Area under the ROC curve (AUC) measures the integral of the two-dimensional area below the ROC curve, thereby yielding an aggregate measure of the classifier's performance across all possible discrimination thresholds (Ferri et al., 2002). AUC is typically interpreted as the probability of the model outperforming random chance. A model with perfect accuracy has an AUC of 1.0 and a model with perfect inaccuracy has an AUC of 0.

The results show that our classifier, fakeRoBERTa, obtains the highest score (AUC = 0.696), followed by OpenAI (AUC = 0.595) and NBSVM (AUC = 0.575). Fig. 5 displays the ROC curves.

Although the results show that fakeRoBERTa outperforms the other classifiers on an independent dataset, this does not necessarily prove the model's generalizability. Looking at raw numbers, fakeRoBERTa's accuracy is 64%, which is 28% better than random chance ((0.64–0.5)/0.5), but still leaves a 36% chance of error. To further increase fakeRoBERTa's generalizability on Ott's dataset, we experimented with fine-tuning. In this process, we infused 80% of Ott's dataset into our Fake Reviews dataset and retrained the classifier using the previously applied fine-tuning approach.

The results from this approach show clear improvement on generalizability, without hindering the performance on the original validation dataset. More specifically, the accuracy on our original CG/OR validation dataset 96.36%, which is nearly identical to the previously recorded performance of 96.91% (a marginal decrease of 0.57%). The accuracy on Ott's dataset is 76.88%, which is more than a 10-percentage point increase from the previous result and decreases the margin of error to 23.12%.

This result can be further improved by using two tactics. First, by (a) *finding the optimal classification threshold* – i.e., the prediction score value that best divides the two classes (Zou et al., 2016). By default, this is

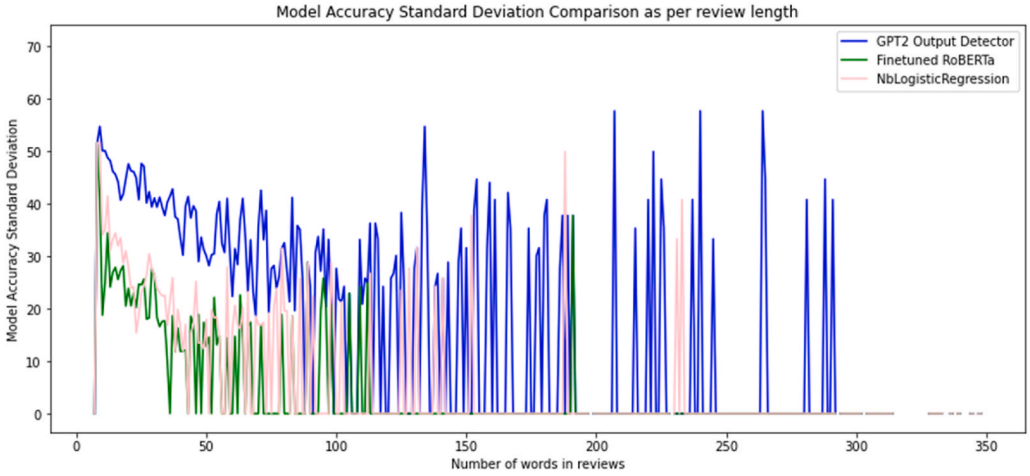

Fig. 4. Standard deviation of predictive accuracy by review length. The graph visualizes accuracy SDs as a line graph, with each model having its own line. The line closest to zero has the smallest variation in accuracy. For the OpenAI model ("GPT-2 output detector model"), we can see substantial variation in performance. The stability of fakeRoBERTa ("Finetuned RoBERTa") and NBSVM ("NBLogisticRegression") is much better. Overall, the performance improves with the length of the review.

Table 5
Variable importance analysis, showing the Top-10 indicator features that support the model's prediction for each class.

Original review		Computer-generated review	
Weight	Feature	Weight	Feature
+9.587	. i	−4.846	this
+9.316	. the	−3.667	. it
+8.385	is a	−3.242	even though
+6.793	have a	−3.083	great
+6.757	!	−3.056	and
+6.511	has the	−2.941	:
+6.367	from	−2.885	it
+6.246	at	−2.678	you
+6.073	and the	−2.602	well
+6.068	, and	−2.560	very

typically 0.5, meaning that if the classifier gives a review a score above 0.5, it is considered fake (positive class), whereas if the score is lower than 0.5, the review is considered real (negative class). To find the optimal classification threshold, we use Youden's J statistic (*J*) (Youden, 1950), which is calculated as the difference between true positive rate (TPR) and false positive rate (FPR):

$$J = \text{TPR} - \text{FPR}$$

The metric is computed for all points in the classifier's ROC curve, and the classification threshold value associated with the maximum value of *J* (i.e., the maximum difference between the true positive rate and the false positive rate) is considered optimal (Schisterman et al., 2005). In our case, we find that the optimal classification threshold is 0.438.

Second, by (b) *increasing the number of epochs* – i.e., the number of times the algorithm processes the training dataset. Increasing the number of epochs may risk overfitting (Li et al., 2019) – a case where the model becomes too adapted on the training set. To avoid this, we examine the error rates in training and validation sets. The results in Fig. 6 indicate that the optimal number of epochs is two, after which the error rates diverge.

Applying these optimizations (classification threshold = 0.438, number of epochs = 2), our classifier reaches an accuracy of 87.81% on Ott and colleagues' dataset. In other words, the error rate is now 12.19%, a major decrease from the previous number. Interestingly, accuracy increases also on our original validation dataset, now being 98.13% (previously 96.36%). All in all, these experiments indicate that, with some fairly straightforward optimization, fakeRoBERTa is also able to detect human-generated fake reviews.

7. Agreement between man and machine

7.1. Hypotheses

We proceeded to ask human annotators to detect if a review was computer-generated or not. This way, we can compare if our classifier is better or worse than people at detecting fake reviews and if the generator model can actually generate reviews that are good enough to fool human reviewers. To this end, we formulate three hypotheses (H):

- **H1:** Machines are better at detecting fake reviews than humans.
- **H2:** Machines agree with each other more than they agree with people.
- **H3:** People agree less with each other than the machines do.

Our premises are that, since fake reviews have detectable but nuanced patterns, it seems reasonable that machines would be better at this task than people and would agree with other machines more often than people. To test these hypotheses, we compute the agreement rates between various classifiers and human raters using Fleiss' Kappa (McHugh, 2012), which is a metric that considers chance-agreement among multiple raters. The procedures for data collection and analysis are explained in the following subsections.

7.2. Data collection

For this experiment, crowd workers annotated a subset of the ML test set that was also used to test the classifier performance. Therefore, the ground-truth values are known to us — i.e., we know if a given review was human-written or generated by our model. Thus, the accuracy of the ML classifiers can be directly compared to that of the human raters. For this purpose, we extracted a random sample of 1000 reviews from the fake reviews dataset. Due to the balanced distribution of the dataset, an equal number of computer- and human-generated reviews was obtained. Overall, this sample size was considered adequate to address the hypotheses.

We recruited crowd workers to assess if a given review was written by a human or generated by a computer. Each review was rated by three crowd workers, resulting in a total of 3000 ratings. The data collection was carried out using the Appen crowdsourcing platform, for which we had an institutional license. A higher quality setting was enabled in the platform, which, according to the platform, results in a smaller group of more experienced, higher-accuracy contributors, based on their historical performance in other tasks on the platform.

The following guidance was given to the crowd workers: "*You will be shown product reviews. Some of them are written by humans, some by*

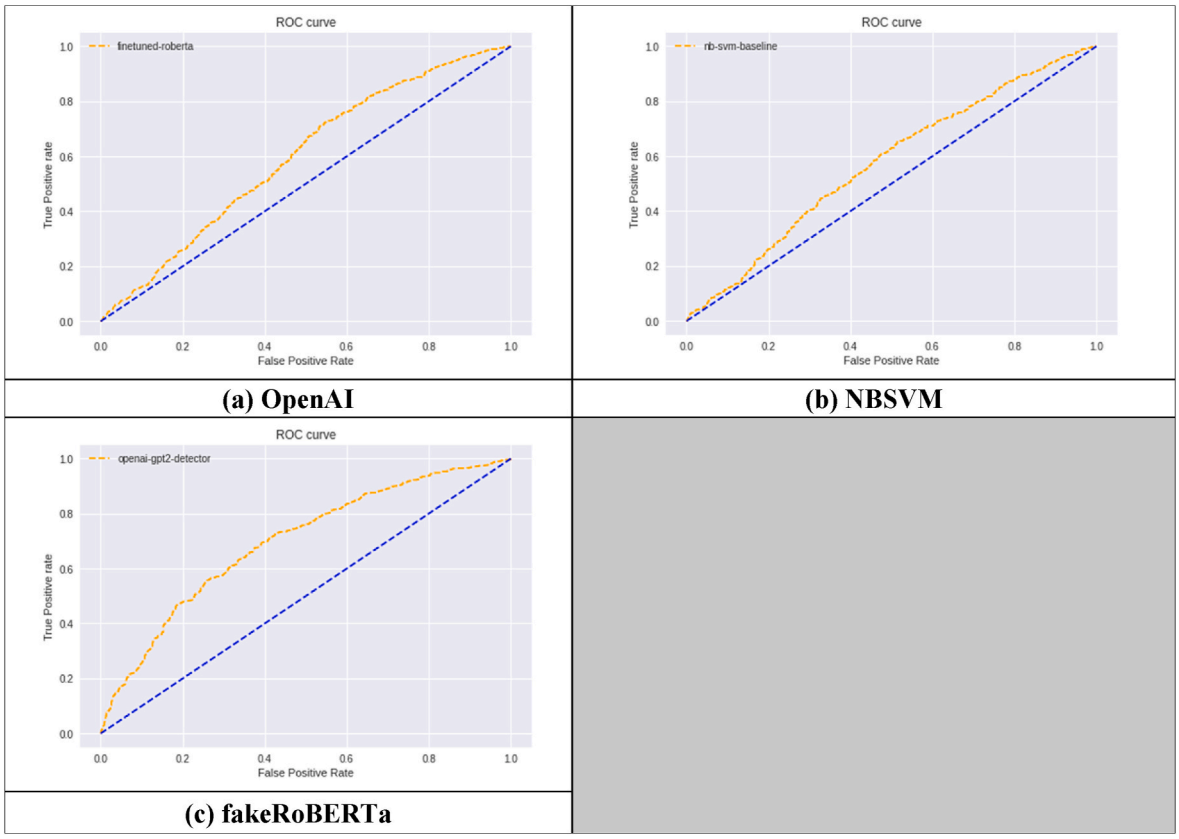

Fig. 5. ROC curves of the classifier on Ott's dataset.

Fig. 6. Train and validation loss. Beyond 2 epochs, the model starts overfitting.

computer. A review created by a human may appear authentic and real. A review created by a computer may appear suspicious or has odd use of language. Give your best assessment of whether a review is written by human or by computer." After this, the crowd workers were shown reviews and asked to indicate their opinion on the following question: "*Do you think this review is generated by a computer or by a human?*" (the answering options were "Computer" or "Human"). One crowd worker could annotate a maximum of 100 reviews, ensuring a minimum of 30 different participants. The workers had to spend at least 3 s viewing a review (on average) to be accepted.

As each review was labeled by three crowd workers, it is possible to always obtain a majority vote because we are dealing with a binary classification task (Alonso, 2015). In other words, with two classes and three raters, one class will always have one vote, and the other will have two votes; hence, the latter will be the most probable candidate. A similar approach of majority voting for fake review detection was previously applied by Harris (2019). Consequently, we obtain this "crowd majority vote" for each of the 1000 reviews and proceed with the statistical analysis of the results.

7.3. Analytical procedure and metrics

Statistical analysis was performed using the statistical software *R* (version 3.6.3). Two-by-two tables were used to evaluate the predictive ability of the majority vote and the three ML classifiers. In order to compare the performance metrics, logistic regression models were fitted for each of the four classifiers. Ten-fold cross-validation was used to construct the 95% confidence interval for accuracy. Since models were fitted on the same versions of the training data, inferences were made to assess the differences between models; this reduces a possible within-resample correlation. Differences were computed using a *t*-test to evaluate the null hypothesis that there is no difference between models.

Accuracy, the Kappa metric, sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV), and AUC were evaluated for each model using original review (OR) as the positive category. The majority rule was used to obtain the final label for human-based models. Fleiss' Kappa was used to assess the inter-rater reliability between the three machine learning models (H2) and three raters (H3). This metric indicates the degree of agreement in classification over that which would be expected by chance, and it can be used to assess the agreement between two or more raters, thus matching our analytical goals.

Statistical analysis using the Z-score, was also performed to assess whether the observed agreement is significantly different from what is expected by chance (under the assumption of no agreement). Percentage agreement was also calculated. Hypothesis testing was performed at a 5% level of significance. A statistically significant result indicates that the agreement between raters is significantly better than would be expected by chance. The p-value, however, does not tell whether the agreement has a high predictive value. The cut-off values in Table 6

Table 6
Interpretation of Fleiss' Kappa (Richard Landis and Koch, 1977).

Value	Interpretation
0	Poor agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1.00	Almost perfect agreement

(Richard Landis and Koch, 1977) were used to interpret the results for Fleiss' Kappa.

8. Results

Results in Table 7 indicate that the fakeRoBERTa model was the highest-performing model among all four prediction sources (accuracy = 96.64%, and Kappa = 0.933), thus corroborating the previous results on the whole test set. The human-based crowd judgment model performed drastically lower than the ML models (accuracy = 55.36%, and Kappa = 0.094). The OpenAI model performed significantly lower than the other two ML models (accuracy = 83% and Kappa = 0.662). A particularly interesting finding is that the crowd judgment is only slightly better than a random guess. Given that there are two options to choose from, the random chance of choosing the correct label is 1/2 = 50%. The crowd judgment is only (55.36–50)/50 ≈ 10.7% better than random chance, whereas even the worst ML (OpenAI) is 66% better than a random guess. As such, H1 is supported: *Machines are better than humans at detecting fake reviews.*

Results in Table 8 show that the overall Fleiss' Kappa for the ML models was $k = 0.717$, which indicates substantial agreement between the ML algorithms. The Kappa coefficient was statistically significant at the 0.1% level, indicating that the probability of observing such agreement by chance is less than 0.1%. The overall percentage agreement among the ML models was 85.9%. The overall Fleiss Kappa for human raters was $k = 0.03$, which indicates poor agreement between raters. The overall percentage agreement among humans was 57.1%, which is only slightly above random chance (14.2%, in precise terms). Therefore, both H2 and H3 are supported: *Machines agree with each other more than they agree with people. People agree less with each other than the machines do.*

Furthermore, the performance of the human-based model was significantly lower compared to the remaining three models ($P < 0.001$), both for accuracy and Kappa (Table 9). No statistically significant difference was observed between the fakeRoBERTa and NBSVM models ($P = 1$), but both models performed significantly better than the OpenAI model and human crowd workers.

9. Discussion and implications

9.1. General discussion

Previous research has not applied transformers for fake-review generation, likely because these technologies were not available before. Therefore, the dataset we release forms an important contribution. Computer-generated reviews have the advantage of being undeniably fabricated – the reviews did not exist before their creation by the algorithm. This is a considerable advantage compared to methods that use statistical analysis (e.g., outlier detection) to determine whether a pre-existing review is fake or not. For example, say that a Person A dislikes a product that most other people like; now, Person A's review may appear as a statistical outlier and might thus be flagged as a fake review, even though it represents a completely truthful opinion. When using unsupervised approaches to annotate datasets for fake review detection, problems such as this may take place.

Our results indicate that human accuracy for detecting fake reviews is only slightly higher than random chance. In other words, the generator can fool humans. Since human accuracy is derived from labels obtained using the majority-vote principle, it appears that the "wisdom of crowds" — i.e., the tendency of accuracy to improve with the participation of more humans in a task (Marbach et al., 2012) — does not improve the accuracy, i.e., humans are not collectively able to form a consensus of a fake review, at least when working independently. Furthermore, the results show much lower agreement among humans than among the ML models, which implies, one the one hand, that people differ by their ability to detect fake reviews and, on the other hand, that machine classifiers perform, regardless of the classifier, more consistently than humans for this detection task. Finally, our results indicate that when applying text-based fake-review detection, the more words a review has, the higher the chance of detecting its true label (fake or real). Longer reviews (>100 words) are easier for the machines to classify correctly than shorter ones (<100 words). This is likely because longer reviews contain more information for a classification decision.

9.2. Theoretical discussion

Computational methods can serve many marketing goals, such as automatic content tagging (Salminen et al., 2019), monitoring of corporate reputation online (Rantanen et al., 2019), optimal budget allocation (Yang et al., 2021), and so on. Most typically, these approaches deal with *positive* effects, i.e., how to create value with ML and AI. However, the other side of the coin is equally important – how to curb the *value destructive* (Makkonen and Olkkonen, 2017) capabilities of technology? (Also known as the "dark side of social media" (Salo et al., 2018)). Our inquiry is situated in this domain, exploring the specific case of fake reviews, and outlining their potential negative impact on the development of electronic marketplaces and the online business sector as a whole. Our study creates awareness of these low-cost, automated methods for creating fake reviews at scale, adaptations of which could be employed by unethical operators in reality. Although we are not the first to reveal the risk of computer-generated reviews (Sun et al., 2013), our findings put forward the need for defense against this type of fraudulent activity.

Regarding the role of humans in the problem of fake reviews, there

Table 8
Agreement between raters (Inter-rater reliability).

Value	Kappa	95% CI	p	Simple agreement
ML algorithms (all)	0.717	0.685–0.75	<0.001	85.9%
NBSVM and fakeRoBERTa only	0.874	0.845–0.903	<0.001	93.7%
Human raters	0.03	0–0.065	0.08	57.1%

95% CI: 95% confidence interval.

Table 7
Performance metrics for human-based and ML models. Highest values highlighted.

	Accuracy	Kappa	Sensitivity	Specificity	PPV	NPV	AUC
Crowd	55.36	0.094	54.71	57.14	77.78	31.52	54.66
NBSVM	95.82	0.916	94.53	97.28	97.53	94.0	95.76
fakeRoBERTa	96.64	0.933	96.17	97.15	97.35	95.87	96.62
OpenAI	83.00	0.662	92.41	76.53	73.02	93.62	83.33

Table 9
Comparison of performance metrics between models.

	Accuracy				Kappa			
	Crowd	NBSVM	fakeRoBERTa	OpenAI	Crowd	NBSVM	fakeRoBERTa	OpenAI
Crowd		−0.4	−0.41	−0.28		−0.82	−0.84	−0.57
NBSVM	**< 0.001**		−0.01	0.13	**< 0.001**		−0.02	0.25
fakeRoBERTa	**< 0.001**	**1**		0.14	**< 0.001**	**1**		0.27
OpenAI	**< 0.001**	**< 0.001**	**< 0.001**		**< 0.001**	**< 0.001**	**< 0.001**	

p-value adjustment: Bonferroni.
Upper diagonal: estimates of the difference.
Lower diagonal: p-value for H0: difference = 0.

can be seen two opposing camps. The first camp argues that human performance is the baseline to beat, i.e., that human knowledge is superior to that of machines (Algur et al., 2010). For example, "The biggest challenge is the lack of an efficient way to tell the difference between a real review and a fake one; even humans are often unable to tell the difference." (p. 1) (Ahmed et al., 2018). The notion here is that 'even humans struggle, therefore how can machines do better?'. In contrast, the other camp argues that humans are not able to detect fake reviews as well as ML models; for example, Ott et al. (2011) and Sun et al. (2013) show this with empirical results. Our findings are aligned with this latter camp, showing that humans are barely able to outperform random chance when assessing if a review was real or not.

In contrast, we show that modern NLP algorithms can effectively detect fake and real reviews with nearly perfect results. The key to this performance, we believe, is that the dataset was large enough (n = 40,000) and balanced, so it affords a good starting point for class separation. These properties made it possible for the algorithms to discover patterns that are not visible to the naked eye. Another explanation for the ML classifiers' superior performance is that the generator model may become biased during its fine-tuning process, which results in repetitive use of certain expressions. These expressions can then be efficiently detected even by standard ML classifiers like the SVM, given there is an adequate amount of training data. This could explain why the ML classifiers' performance so drastically overshadows that of humans – due to cognitive limitations, humans are unable to process the vast amount of training data to identify such patterns. Therefore, due to cognitive limitations of humans to detect patterns from unstructured data (Kahneman and Amos, 1972, 1973), detecting fake reviews at scale may already be beyond human capability.

Overall, the results suggest that it is becoming increasingly difficult for consumers to distinguish high-quality products from low-quality ones based on online reviews. This is a major debacle that can be considered one of the most prominent risks for e-commerce, mainly because *trust* has an elevated role in electronic marketplaces (Papadopoulou et al., 2001; Tran and Strutton, 2020; Jacob et al., 2011). At its worst, the proliferation of fake content results in adverse selection (Akerlof, 1970), according to which consumers lose their ability to detect good products from bad ones. The greatest strength of eWOM is that consumer-to-consumer (C2C) reviews are perceived as trustworthy, credible, and less biased than company-generated information (Costa et al., 2019), thus providing a useful signal of a product's true quality (Akerlof, 1970). If this signal is distorted, the ramifications to e-commerce and other forms of online business would be tremendous. Therefore, as the theory holds, fake reviews represent a type of fraudulent activity that can have far-reaching negative implications for entire industries.

Theoretically, fake review detection can be understood as a "cat and mouse game," where generators and detectors compete against one another. The goal of the generator is to create reviews that are undetectable by both humans and machines; the goal of the detector is to prevent this by always identifying signals of fabrication. As the detection methods evolve, malicious users develop more advanced camouflage strategies in response (Wu et al., 2017), mitigating the performance of detectors over time. For example, to avoid detection, fake review writers may consciously use words and expressions that appear authentic and avoid reusing the same words in a repetitive way (Mukherjee et al., 2013). This renders methods relying on simple techniques such as duplicate detection (Algur et al., 2010; Jindal and Liu, 2008) basically useless, as fake reviewers circumvent them by introducing unforeseen variation. The only plausible solution is to keep developing datasets and methods for fake review detection. One interesting approach is online learning (Cardoso et al., 2018), in which continuous feedback improves performance over time.

9.3. Practical implications

The study findings provide implications for three stakeholder groups.

(i) *Firms* are advised to **refrain from participating in fake review generation** (apart from possible experimental purposes). An example of fake reviews is the case of Samsung. Taiwan's Fair Trade Commission showed that Samsung's Taiwanese unit had recruited writers to create online reviews presenting Samsung's smartphones in a positive light and highlighting flaws in a competitor's products.[4] This practice was seen as violating laws of fair trade and resulted in Samsung being fined. Hence, engaging in fraudulent review practices poses a major (financial and reputational) risk for the violating company. Nonetheless, firms may also find commercial opportunities emerging around the fake review detection industry, including **developing new services for fake review detection.** For example, Fakespot (fakespot.com) is specialized in the detection of fake Amazon product reviews. Similar new services may be developed, e.g., to **monitor the health of a company's online reviews on a consistent basis**, which is a practice that companies should adopt to detect potential review attacks. Without a proper response, fake reviews can turn into a brand reputation crisis if they are not removed.

(ii) *Consumers* face the dilemma of trust concerning online reviews. On the one hand, reviews are extremely helpful, as they provide vital information for people to spend their hard-earned money on products and services with satisfactory quality. While online reviews are useful for this goal, blind trust in them is perilous for consumers. Gaining awareness of fake reviews is quintessential for understanding the risks of using online reviews in purchase decisions, and consumers need to be alert that in the current environment, reviews may not even be written by humans but instead generated by computers. As some proportion of online reviews is faked for profit, decisions based on online reviews should be made carefully – e.g., by **reading and comparing review texts** instead of merely relying on aggregate scores and

[4] https://www.smh.com.au/technology/samsung-fined-for-hiring-bloggers-to-write-fake-reviews-attack-rival-htc-20131025-2w5nx.html

ratings. Various online resources[5] exist to help consumers assess if a review is fake; consumers are therefore encouraged to **get educated about fake reviews by using research-based resources.**

(iii) *Platforms* offering reviewing possibilities (such as Google, Facebook, Amazon, Hotels.com, TripAdvisor, Yelp …) are encouraged to **improve agility of reacting to fake reviews**, which is critical for protecting the trustworthiness of these platforms and maintaining a high-quality user experience for consumers seeking information in these platforms. Platforms should **enable users an easy way to report fraudulent activity**, which requires user interface research (Su'a et al., 2017). In terms of managing the error given by classifiers, the platforms need to **ensure adequate human resources to investigate machine-flagged reviews** using contextual information, such as the reviewer's profile information and previous activity. Manual verification is important because automatically removing machine-flagged reviews would violate users' rights in cases where the algorithm misclassifies a truthful review as a fake one. Finally, platforms should **provide a way to reverse positive and negative effects due to fraudulent reviews** – such that reputational damage due to fake reviews would not remain permanent and the better positions obtained by fake reviews would be lowered in due course by ranking algorithms. It is in the platforms' interest to prevent loss of consumer trust, as their business models ultimately depend on this trust.

The ramifications of fake reviews extend beyond consumer trust. Namely, commercial systems such as sentiment analysis and online opinion mining (Cambria et al., 2013) can be biased by a large number of junk reviews. Consider, for example, a firm that monitors its social media reputation by using automated tools that mine reviews of the firm's products. Now, if fake reviews were to bias the sentiment distribution, the firm would obtain a misleading score concerning its online reputation. Therefore, the practical implications of fake reviews are not contained to consumer-facing platforms alone but also apply to downstream applications such as sentiment analysis and opinion mining systems.

9.4. Limitations and future research

The study contains some limitations, which we discuss here.

First, there are technical aspects in text generation and classification that could benefit from future experiments. For example, we could create benchmarks of classification performance for different sampling techniques used for text generation. In the current study, we focused primarily on top-k nucleus sampling to generate our reviews, but different sampling techniques could be investigated.

Second, one limitation is that the original Amazon review dataset might involve an unknown number of fake reviews already. If so, this would result in a bias for the language model we applied. Unfortunately, we have no way of knowing whether a given review in the dataset is undeniably truthful. As a practical solution, we assumed that this number is low (<5%), in which case it does not bias the generator in any significant way. The only way to utilize real reviews is to make such assumptions, as there is no way to know the precise motivations of each user writing a review. One factor that alleviates this risk is that Amazon already employs fake review detection; He et al. (2021) observed that a large number of reviews they were tracking were deleted by Amazon over time.

Third, a major impediment to globally generalizable fake review detectors is the fact that most of the known datasets (including ours) are in English (an exception is a dissertation by Abu Hammad (2013) that develops a classifier for the Arabic language). In a similar vein, the best generator models are available for English. Therefore, there is a substantial lack of research and development in other languages such as Arabic, Hindi, German, French, Spanish, Russian, and so on. The e-commerce sector and other types of online businesses extend beyond borders, cultures, and languages, and therefore there is no reason to only focus on one language, even if English is the current *lingua franca*. Technically, algorithms need to be trained to reflect each language to increase global applicability.

Fourth, a general limitation of any current text generators is that they have no awareness, motives, or understanding of what they are doing. Their understanding of language is based on mimicry — on millions of examples of how people communicate online, which is then fine-tuned for the context of language used in online reviews. Therefore, even though the reviews may sometimes appear as if they express original thoughts, this is merely the side-effect of mimicry, not actual creative thinking. We mention this general limitation of current AI/ML to give the reader a realistic picture of the state of technology, and to avoid the hyperbolic statements attached to AI/ML (Oravec, 2019). AI/ML is "artificial" (mimicked) intelligence, instead of human intelligence that includes self- and social awareness. This property also makes it more difficult to control the precise sentiment and aspects in the generated reviews, meaning a human writer can be given a precise instruction to "write a review about Product X that praises the product's weight and color," but asking a generator to do the same is much more difficult. To this end, it might also be interesting to evaluate a combination of machine-generated fake reviews that are then edited by humans to disrupt grammar, linguistic, and spacing patterns that the machine learns.

Finally, the results indicate that fakeRoBERTa can successfully predict both computer- and human-generated reviews. However, ML models face the general caveat of dataset specificity, so as the nature of human-generated reviews evolves over time, the only way to maintain a high performance is frequently updating the classifiers. Therefore, future work needs to pursue the creation of trustworthy baseline datasets, especially of large human-generated fake reviews. Also, because the nature of communication (e.g., micro-text reviews in Twitter versus longer reviews in Amazon) differs by platform, the applicability of fake detection classifiers across platforms should be examined. Again, this requires that not only e-commerce product reviews but also other forms of reviews taking place in social media are included in fake review datasets.

9.5. Concluding remarks

Earlier in this section, we described fake review detection as a game of cat and mouse. This analogy technically parallels one important development in deep learning, namely generative adversarial networks (GANs) (Goodfellow et al., 2014). The purpose of GANs is precisely the same – a generator network takes an input of randomized samples and iteratively modifies the outputs based on a discriminator network's feedback. Through this process, the outputs gradually become more realistic until the point where the discriminator can no longer distinguish real from artificial. As far as we know, GANs have not been applied for the purpose of fake review detection thus far. Therefore, this particular technology seems like the next logical step to experiment with.

To support further development in this area, we make our *Fake Review Dataset* (v 1.0) publicly available.[6] An advantage of this dataset is that it is equally balanced between positive and negative cases. As Crawford et al. (2015) note, many pre-existing datasets in this domain are highly imbalanced due to the fact that truthful reviews tend to be much more frequent than fake ones. Imbalanced datasets require

[5] https://www.consumerprotectionbc.ca/2017/01/can-you-spot-a-fake-online-review/

[6] https://osf.io/tyue9/

elaborate algorithmic considerations (Al Najada and Zhu, 2014) that may or may not be successful. These specificities are not required for the development of unbiased classification models when using our dataset.

Furthermore, we share all the test datasets, model parameters, model weights, and initialization seeds for experimentation reproducibility.[7] Unfortunately, sharing results for replication has thus far been too rare in this field. We hope that our example of sharing the results and computational notebooks serves as a positive example for changing this. Similarly, the lack of gold standard datasets and predefined performance benchmarks represents a joint challenge for the field to overcome (Viviani and Pasi, 2017). As a broader implication, sharing resources is a crucial step toward empirical AI research in marketing (Mustak et al., 2021) – without this, every study has to start from the bare beginning.

Finally, an interesting and important question is, *what does fake mean for marketing?* With malicious actors increasingly using online platforms to spread misinformation in the form of synthetically created images, videos, audio, and texts, how should marketers react? These questions go beyond the scope of this manuscript, but we nonetheless highlight the need for a stream of studies investigating these questions, including how to create better technical and non-technical countermeasures to combat fake creations. In this effort, technology is both an ally and an adversary. While text-generation algorithms could be employed by evil marketers for large-scale production of fake reviews, they can also be used by ethical marketers to develop countermeasures such as more efficient detectors to deter deception.

10. Conclusion

Detection of fake reviews is a problem for researchers, e-commerce sites, and firms engaged in online business. Our results indicate that current text generation methods yield fake reviews that appear so realistic that it is challenging for a human to detect them. Fortunately, machine learning classifiers do much better in this regard, with almost perfect accuracy in detecting reviews generated by other machines. This implies that "machines can fight machines" in the battle against fake reviews. Future research is needed for experimenting with more datasets and platforms.

References

Abu Hammad, Ahmad Sj, 2013. An Approach for Detecting Spam in Arabic Opinion Reviews. Doctoral Dissertation. Islamic University of Gaza, Gaza, Palestine.

Ahmed, Hadeer, Traore, Issa, Saad, Sherif, 2018. Detecting opinion spams and fake news using text classification. e9 Security and Privacy 1, 1, 2018.

Akerlof, George A., 1970. The market for "lemons": quality uncertainty and the market mechanism. Q. J. Econ. 84 (3), 488–500. https://doi.org/10.2307/1879431. August 1970.

Al Najada, Hamzah, Zhu, Xingquan, 2014. iSRD: spam review detection with imbalanced data distributions. In: Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014). IEEE, pp. 553–560.

Algur, Siddu P., Patil, Amit P., Hiremath, P.S., Shivashankar, S., 2010. Conceptual level similarity measure based review spam detection. In: 2010 International Conference on Signal and Image Processing. IEEE, pp. 416–423.

Alonso, Omar, 2015. Practical lessons for gathering quality labels at scale. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15). ACM, New York, NY, USA, pp. 1089–1092. https://doi.org/10.1145/2766462.2776778.

Bell, Jason, 2014. In: Machine Learning: Hands-On for Developers and Technical Professionals, first ed. John Wiley & Sons, Hoboken, New Jersey.

Bradley, Andrew P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 30 (7), 1145–1159, 1997.

Budzianowski, Pawel, Vulić, Ivan, 2019. Hello, it's GPT-2–how can I help you? towards the use of pretrained language models for task-oriented dialogue systems, 2019 arXiv. preprint arXiv:1907.05774.

Cambria, E., Schuller, B., Xia, Y., Havasi, C., 2013. New avenues in opinion mining and sentiment analysis. IEEE Intell. Syst. 28 (2), 15–21. https://doi.org/10.1109/MIS.2013.30. March 2013.

Cardoso, Emerson F., Silva, Renato M., Almeida, Tiago A., 2018. Towards automatic filtering of fake reviews. Neurocomputing 309, 106–116. https://doi.org/10.1016/j.neucom.2018.04.074. October 2018.

Clarkson, Philip R., Robinson, Anthony J., 1997. Language model adaptation using mixtures and an exponentially decaying cache. In: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, pp. 799–802.

Costa, Ana, Guerreiro, João, Moro, Sérgio, Henriques, Roberto, 2019. Unfolding the characteristics of incentivized online reviews. J. Retailing Consum. Serv. 47, 272–281, 2019.

Crawford, Michael, Khoshgoftaar, Taghi M., Prusa, Joseph D., Richter, Aaron N., Al Najada, Hamzah, 2015. Survey of review spam detection using machine learning techniques. Journal of Big Data 2 (1), 23. https://doi.org/10.1186/s40537-015-0029-9. October 2015.

DePaulo, Bella M., Kashy, Deborah A., Kirkendol, Susan E., Wyer, Melissa M., Epstein, Jennifer A., 1996. Lying in everyday life. J. Pers. Soc. Psychol. 70 (5), 979, 1996.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, Toutanova, Kristina, 2018. Bert: pre-training of deep bidirectional transformers for language understanding arXiv preprint arXiv:1810.04805 (2018).

Duarte, Paulo, e Silva, Susana Costa, Ferreira, Margarida Bernardo, 2018. How convenient is it? Delivering online shopping convenience to enhance customer satisfaction and encourage e-WOM. J. Retailing Consum. Serv. 44, 161–169, 2018.

Dwivedi, Yogesh K., Ismagilova, Elvira, Laurie Hughes, D., Carlson, Jamie, Filieri, Raffaele, Jacobson, Jenna, Jain, Varsha, Karjaluoto, Heikki, Kefi, Hajer, Krishen, Anjala S., Kumar, Vikram, Rahman, Mohammad M., Raman, Ramakrishnan, Rauschnabel, Philipp A., Rowley, Jennifer, Salo, Jari, Tran, Gina A., Wang, Yichuan, 2020. Setting the future of digital and social media marketing research: perspectives and research propositions. Int. J. Inf. Manag. 102168. https://doi.org/10.1016/j.ijinfomgt.2020.102168. July 2020.

Endo, Seiji, Yang, Jun, Park, JungKun, 2012. The investigation on dimensions of e-satisfaction for online shoes retailing. J. Retailing Consum. Serv. 19 (4), 398–405, 2012.

Ferri, César, Flach, Peter, Hernández-Orallo, José, 2002. Learning decision trees using the area under the ROC curve. ICML 139–146.

Filieri, Raffaele, 2016. What makes an online consumer review trustworthy? Ann. Tourism Res. 58 (May 2016), 46–64. https://doi.org/10.1016/j.annals.2015.12.019.

Floridi, Luciano, 2018. Artificial intelligence, deepfakes and a future of ectypes. Philosophy & Technology 31 (3), 317–321, 2018.

Floridi, Luciano, Chiriatti, Massimo, 2020. GPT-3: its nature, scope, limits, and consequences. Minds Mach. 1–14, 2020.

Fontanarava, Julien, Gabriella, Pasi, Viviani, Marco, 2017. Feature analysis for fake review detection through supervised classification. In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA). IEEE, pp. 658–666.

François, Thomas, Miltsakaki, Eleni, 2012. Do NLP and machine learning improve traditional readability formulas? Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations 49–57.

Gobi, N., Rathinavelu, A., 2019. Analyzing cloud based reviews for product ranking using feature based clustering algorithm. Cluster Comput. 22 (3), 6977–6984, 2019.

Goodfellow, Ian J., Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, Bengio, Yoshua, 2014. Generative Adversarial Networks arXiv:1406.2661 [cs, stat] (June 2014). Retrieved February 27, 2018 from. http://arxiv.org/abs/1406.2661.

Hajek, Petr, Henriques, Roberto, 2017. Mining corporate annual reports for intelligent detection of financial statement fraud – a comparative study of machine learning methods. Knowl. Base Syst. 128 (July 2017), 139–152. https://doi.org/10.1016/j.knosys.2017.05.001.

Harris, Christopher G., 2019. Comparing human computation, machine, and hybrid methods for detecting hotel review spam. In: Digital Transformation for a Sustainable Society in the 21st Century (Lecture Notes in Computer Science). Springer International Publishing, Cham, pp. 75–86. https://doi.org/10.1007/978-3-030-29374-1_7.

He, Sherry, Hollenbeck, Brett, Proserpio, Davide, 2021. The Market for Fake Reviews. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3664992.

Howard, Jeremy, Ruder, Sebastian, 2018. Universal Language Model Fine-Tuning for Text Classification arXiv preprint arXiv:1801.06146 (2018).

Ismagilova, Elvira, Emma Slade, Rana, Nripendra P., Dwivedi, Yogesh K., 2020. The effect of characteristics of source credibility on consumer behaviour: a meta-analysis. J. Retailing Consum. Serv. 53 (March 2020), 101736. https://doi.org/10.1016/j.jretconser.2019.01.005.

Jacob, Weisberg, Te'eni, Dov, Arman, Limor, 2011. Past Purchase and Intention to Purchase in E-Commerce: the Mediation of Social Presence and Trust. Internet research, 2011.

Jan, Kietzmann, Lee, Linda W., McCarthy, Ian P., Kietzmann, Tim C., 2020. Deepfakes: trick or treat? Bus. Horiz. 63 (2), 135–146, 2020.

Jindal, Nitin, Liu, Bing, 2008. Opinion spam and analysis. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 219–230.

Kaabachi, Souheila, Ben Mrad, Selima, Petrescu, Maria, 2017. Consumer initial trust toward internet-only banks in France. Int. J. Bank Market. 35 (6), 903–924. https://doi.org/10.1108/IJBM-09-2016-0140. January 2017.

Kahneman, Daniel, Amos, Tversky, 1972. Subjective probability: a judgment of representativeness. Cognit. Psychol. 3 (3), 430–454, 1972.

Kahneman, Daniel, Amos, Tversky, 1973. On the psychology of prediction. Psychol. Rev. 80 (4), 237, 1973.

[7] https://drive.google.com/drive/folders/1aadnypFL_aLiP6Z61VAW4_gYibGXLV2x?usp=sharing

Kaushik, Kapil, Mishra, Rajhans, Rana, Nripendra P., Dwivedi, Yogesh K., 2018. Exploring reviews and review sequences on e-commerce platform: a study of helpful reviews on Amazon. J. Retailing Consum. Serv. 45, 21–32, 2018.

Kirkpatrick, Keith, 2016. Battling algorithmic bias: how do we ensure algorithms treat us fairly? Commun. ACM 59 (10), 16–17, 2016.

Lee, Kyungyup Daniel, Han, Kyungah, Myaeng, Sung-Hyon, 2016. Capturing word choice patterns with LDA for fake review detection in sentiment analysis. In: Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics (WIMS '16), 1–7. Association for Computing Machinery, New York, NY, USA.

Li, Haidong, Li, Jiongcheng, Guan, Xiaoming, Liang, Binghao, Lai, Yuting, Luo, Xinglong, 2019. Research on overfitting of deep learning. In: 2019 15th International Conference on Computational Intelligence and Security (CIS). IEEE, pp. 78–81.

Liang, Yuding, Zhu, Kenny, 2018. Automatic generation of text descriptive comments for code blocks. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, Veselin Stoyanov, 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach arXiv:1907.11692 [cs] (July 2019). Retrieved April 26, 2021 from. http://arxiv.org/abs/1907.11692.

Luca, Michael, 2011. Reviews, reputation, and revenue: the case of Yelp.Com. SSRN Journal. https://doi.org/10.2139/ssrn.1928601, 2011.

Makkonen, Hannu, Olkkonen, Rami, 2017. Interactive value formation in interorganizational relationships: dynamic interchange between value co-creation, no-creation, and co-destruction. Market. Theor. 17 (4), 517–535, 2017.

Marbach, Daniel, Costello, James C., Küffner, Robert, Vega, Nicole M., Prill, Robert J., Camacho, Diogo M., Allison, Kyle R., Kellis, Manolis, Collins, James J., Stolovitzky, Gustavo, 2012. Wisdom of crowds for robust gene network inference. Nat. Methods 9 (8), 796–804, 2012.

Mattson, Christopher, Bushardt, Reamer L., Artino Jr., Anthony R., 2021. When a Measure Becomes a Target, it Ceases to Be a Good Measure.

McHugh, Mary L., 2012. Interrater reliability: the kappa statistic. Biochem. Med. 22 (3), 276–282. October 2012.

Mihalcea, Rada, Strapparava, Carlo, 2009. The lie detector: explorations in the automatic recognition of deceptive language. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 309–312.

Mukherjee, Arjun, Venkataraman, Vivek, Liu, Bing, Glance, Natalie, 2013. What yelp fake review filter might be doing?. In: Proceedings of the International AAAI Conference on Web and Social Media.

Munzel, Andreas, 2016. Assisting consumers in detecting fake reviews: the role of identity information disclosure and consensus. J. Retailing Consum. Serv. 32, 96–108, 2016.

Munzel, Andreas, Kunz, Werner H., 2014. Creators, multipliers, and lurkers: who contributes and who benefits at online review sites. Journal of Service Management, 2014.

Mustak, Mekhail, Salminen, Joni, Loïc, Plé, Wirtz, Jochen, 2021. Artificial intelligence in marketing: topic modeling, scientometric analysis, and research agenda. J. Bus. Res. 124 (January 2021), 389–404. https://doi.org/10.1016/j.jbusres.2020.10.044.

Ni, Jianmo, Li, Jiacheng, McAuley, Julian, 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 188–197.

Oravec, Jo Ann, 2019. Artificial intelligence, automation, and social welfare: some ethical and historical perspectives on technological overstatement and hyperbole. Ethics Soc. Welfare 13 (1), 18–32, 2019.

Ott, Myle, Choi, Yejin, Cardie, Claire, Jeffrey, T., Hancock, 2011. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557* (2011).

Papadopoulou, Panagiota, Andreou, Andreas, Kanellis, Panagiotis, Martakos, Drakoulis, 2001. Trust and Relationship Building in Electronic Commerce. Internet research, 2001.

Petrescu, Maria, O'Leary, Kathleen, Goldring, Deborah, Ben Mrad, Selima, 2018. Incentivized reviews: promising the moon for a few stars. J. Retailing Consum. Serv. 41, 288–295, 2018.

Plotkina, Daria, Munzel, Andreas, Pallud, Jessie, 2020. Illusions of truth—experimental insights into human and algorithmic detections of fake online reviews. J. Bus. Res. 109 (March 2020), 511–523. https://doi.org/10.1016/j.jbusres.2018.12.009.

Radford, Alec, Wu, Jeffrey, Child, Rewon, Luan, David, Amodei, Dario, Sutskever, Ilya, 2019. Language models are unsupervised multitask learners. OpenAI blog 1 (8), 9, 2019.

Raffel, Colin, Shazeer, Noam, Roberts, Adam, Lee, Katherine, Narang, Sharan, Matena, Michael, Zhou, Yanqi, Li, Wei, Liu, Peter J., 2019. Exploring the Limits of Transfer Learning with a Unified Text-To-Text Transformer. arXiv preprint arXiv: 1910, p. 10683, 2019.

Rantanen, Anette, Salminen, Joni, Ginter, Filip, Bernard, J., Jansen, 2019. Classifying online corporate reputation with machine learning: a study in the banking domain. Internet Res. 30 (1) https://doi.org/10.1108/INTR-07-2018-0318 (January 2019).

Richard Landis, J., Koch, Gary G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159. https://doi.org/10.2307/2529310 (March 1977).

Salminen, Joni, Almerekhi, Hind, Milenković, Milica, Jung, Soon-gyo, An, Jisun, Kwak, Haewoon, Bernard, J., Jansen, 2018. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In: Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2018), San Francisco, California, USA. Retrieved from. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17885.

Salminen, Joni, Yoganathan, Vignesh, Juan Corporan, Jansen, Bernard J., Jung, Soon-Gyo, 2019. Machine learning approach to auto-tagging online content for content marketing efficiency: a comparative analysis between methods and content type. J. Bus. Res. 101, 203–217. https://doi.org/10.1016/j.jbusres.2019.04.018. August 2019.

Salo, Jari, Mäntymäki, Matti, Islam, AKM Najmul, 2018. The dark side of social media–and Fifty Shades of Grey introduction to the special issue: the dark side of social media. Internet Res. 28, 5. https://doi.org/10.1108/IntR-10-2018-442, 2018.

Sandra Mc Loureiro, Luisa Cavallero, Javier Miranda, Francisco, 2018. Fashion brands on retail websites: customer performance expectancy and e-word-of-mouth. J. Retailing Consum. Serv. 41, 131–141, 2018.

Sandulescu, Vlad, Ester, Martin, 2015. Detecting singleton review spammers using semantic similarity. In: Proceedings of the 24th International Conference on World Wide Web, pp. 971–976.

Schisterman, Enrique F., Perkins, Neil J., Liu, Aiyi, Howard, Bondell, 2005. Optimal cut-point and its corresponding youden index to discriminate individuals using pooled blood samples. Epidemiology 16 (1), 73–81. https://doi.org/10.1097/01.ede.0000147512.81966.ba. January 2005.

Shivagangadhar, Kolli, Sagar, H., Sathyan, Sohan, Vanipriya, C.H., 2015. Fraud detection in online reviews using machine learning techniques. Int. J. Comput. Eng. Res. 5 (5), 52–56, 2015.

Smith, A., Anderson, M., 2016. *Online Shopping and E-Commerce, Online Reviews*. Pew Research Center. Retrieved from. https://www.pewinternet.org/2016/12/19/online-reviews/.

Sun, Huan, Morales, Alex, Yan, Xifeng, 2013. Synthetic review spamming and defense. In: Proceedings of the 22nd International Conference on World Wide Web Companion, 9. Rio de Janeiro, Brazil.

Su'a, Tavita, Licorish, Sherlock A., Roy Savarimuthu, Bastin Tony, Langlotz, Tobias, 2017. Quick review: a novel data-driven mobile user interface for reporting problematic app features. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces, pp. 517–522.

Tolosana, Ruben, Vera-Rodriguez, Ruben, Fierrez, Julian, Morales, Aythami, Ortega-Garcia, Javier, 2020. Deepfakes and beyond: a survey of face manipulation and fake detection. Inf. Fusion 64 (2020), 131–148.

Tran, Gina A., Strutton, David, 2020. Comparing email and SNS users: investigating e-service scape, customer reviews, trust, loyalty and E-WOM. J. Retailing Consum. Serv. 53, 101782, 2020.

Viviani, Marco, Pasi, Gabriella, 2017. Credibility in social media: opinions, news, and health information—a survey. e1209 Wiley interdisciplinary reviews: Data Min. Knowl. Discov. 7, 5, 2017.

Wang, Sida, Manning, Christopher D., 2012. Baselines and bigrams: simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. Association for Computational Linguistics, pp. 90–94.

Wolf, Thomas, Debut, Lysandre, Sanh, Victor, Chaumond, Julien, Clement, Delangue, Anthony, Moi, Cistac, Pierric, Tim Rault, Louf, Rémi, Morgan, Funtowicz, 2019. HuggingFace's Transformers: State-Of-The-Art Natural Language Processing arXiv preprint arXiv:1910.03771 (2019).

Wu, Xian, Dong, Yuxiao, Tao, Jun, Huang, Chao, Nitesh, V., Chawla, 2017. Reliable fake review detection via modeling temporal and behavioral patterns. In: 2017 IEEE International Conference on Big Data. Big Data), pp. 494–499. https://doi.org/10.1109/BigData.2017.8257963.

Yang, Yanwu, Feng, Baozhu, Salminen, Joni, Bernard, J., Jansen, 2021. Optimal advertising for a generalized Vidale–Wolfe response model. Electron. Commer. Res. https://doi.org/10.1007/s10660-021-09468-x. March 2021.

Yoo, Kyung-Hyan, Gretzel, Ulrike, 2009. Comparison of deceptive and truthful travel reviews. In: Information and Communication Technologies in Tourism 2009. Springer, pp. 37–47.

Youden, W.J., 1950. Index for rating diagnostic tests. Cancer 3 (1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3. January 1950.

Zou, Quan, Xie, Sifa, Lin, Ziyu, Wu, Meihong, Ju, Ying, 2016. Finding the best classification threshold in imbalanced classification. Big Data Research 5 (September 2016), 2–8. https://doi.org/10.1016/j.bdr.2015.12.001.