Technological Forecasting & Social Change 189 (2023) 122380

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

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

Efficiency of automatic text generators for online review content generation

A. Perez-Castro [a], M.R. Martínez-Torres [a,*], S.L. Toral [b]

[a] Facultad de Ciencias Económicas y Empresariales, University of Seville (Spain), Av. de Ramón y Cajal, 1, 41018 Sevilla, Spain
[b] E. T. S. Ingeniería, University of Seville (Spain), Avda. Camino de los Descubrimientos s/n, 41092 Seville, Spain

ARTICLE INFO

Keywords:
Deceptive reviews generation
Word-based encoding
Context-based encoding
Pretrained models
Transfer learning

ABSTRACT

The evolution of Artificial Intelligence has led to the appearance of automatic text generators able to closely resemble human writing, endangering the development of e-commerce and the consumer confidence. Thus, it is critical to deeply understand how these text generators work to present the presence of deceptive reviews. This paper analyzes one of the most popular text generators, GPT2 (Generative Pre-trained Transformer 2), and studies its effectivity compared to human-generated reviews using previously published classifiers trained to distinguish between real and deceptive reviews. One parameter of the model is the so-called temperature, which determines how deterministic the model is. The temperature adjusts the probability distribution of the words in the model, so that a higher temperature translates into a higher degree of inventiveness in the generation of the texts. Findings reveal (i) that automatically-generated deceptive reviews worsen the accuracy of existing classifiers, this effect being accentuated by the degree of inventiveness; (ii) that their performance depends on the data used to train the generator; and (iii) that the sentiment polarity has no effect on the performance of detection classifiers.

1. Introduction

Multiple studies have shown that online consumer reviews are a key source of information, affecting not only the purchasing decisions of potential consumers but also the reputation of products and services (Filieri et al., 2015; Petrescu et al., 2018; Zhou et al., 2021). While favorable reviews might result in financial rewards, unfavorable reviews can damage a reputation brand, causing financial loss (Zhu and Zhang, 2010). As a result, the proliferation of fake reviews is becoming more and more common every day. Many businesses have used this method to attract new customers or even to defeat their competitors with fake reviews (Wu et al., 2020).

At this point, it is important to differentiate between two types of fake reviews: those created by humans and those created automatically by machines through bots or text generators. Initially, fake reviews could only be created manually by humans in order to be indistinguishable from a truthful review, but with advances in generative text models, it is now possible to do it artificially by training them with texts that belong to a specific domain type (Köbis and Mossink, 2021). These bots, generally based on language models, excel in the generation of fluent and coherent text, very similar to the human style, making the arduous task of differentiating between real and fake reviews even more difficult than those fake reviews created by humans. Hence, a new generation of fraudulent reviews has become a new challenge with significant consequences for e-commerce (Yao et al., 2017). Moreover, these language models are shared and pre-trained on the Internet, which exposes consumers to an even greater threat, as anyone with some knowledge in computer science can easily build their own fake reviews generator.

Previous work has analyzed the social impact of artificial intelligence in the generation of fake news (Ahmad et al., 2022), the spread of malicious rumors (Meel and Vishwakarma, 2020) or the generation of false images or videos (Karnouskos, 2020) with the aim of devaluing the public image of people, governments or institutions. This impact translates into disinformation, the polarization of society and the dissemination of conspiracy theories without any evidence. These studies focus primarily on the impact, i.e. the consequences. However, it is equally important to understand the mechanisms underlying the generative algorithms of artificial intelligence in order to be able to define appropriate mitigation strategies for their effects.

The aim of this paper is to characterize the efficiency of a fake review generator in terms of its ability to mislead different classifiers of fake reviews widely used in other studies. More specifically, this research will focus on studying the efficiency of one of the most famous text

* Corresponding author.
E-mail addresses: ampiperez97@gmail.com (A. Perez-Castro), rmtorres@us.es (M.R. Martínez-Torres), storal@us.es (S.L. Toral).

https://doi.org/10.1016/j.techfore.2023.122380
Received 4 January 2022; Received in revised form 24 January 2023; Accepted 28 January 2023
Available online 4 February 2023

generators: GPT2 (Generative Pre-trained Transformer 2), a language model based on the transformer architecture created by OpenAI (https://openai.com/). Different experiments will be carried out to compare the performance of the classifiers when detecting automatically-generated deceptive reviews using GTP2 or human-generated deceptive reviews. The aim of the proposed experiments is to evaluate how indistinguishable the automatic and human-generated reviews are. Furthermore, several characteristics of GPT2 will be studied, such as the generation of sentiment-preserving reviews and the possibility of modulating the degree of inventiveness.

The main contributions of this study are:

- The creation of a sentiment-preserving review generator, by modifying the GPT2-model, so that positive/negative reviews of varying length can be generated.
- The analysis of the effectivity of GTP2 for the fake review generation task, considering the effect of sentiment and inventiveness on the final accuracy.
- The study of the performance of classifiers dealing with human and automatically generated fake reviews.

The remainder of the paper is structured as follows: Section 2 details the related work in the field of deceptive review detection and generation, focusing on the different techniques for the generation of fake reviews. Section 3 formulates the research framework, detailing the selected architectures for generation and classification. Section 4 describes the annotated datasets used as case studies, and the setup of the experiments. Section 5 provides the results and comparison of the selected approaches. Section 6 discusses the results and their implications. Finally, Section 7 summarizes the conclusions of this work.

2. Related work

In this section, we review existing work related to the generation of false reviews in general (Section 2.1), and then focus on algorithmic text generation methods, also known as text generators (Section 2.2). This section initially introduces recurrent neural networks (RNN), which is the architecture commonly used for natural language processing, given its ability to find the temporal dependencies needed to generate syntactically and grammatically coherent text. Next, we introduce the three main deep-generative methods currently used and based on RNN that allow the training dataset to be modelled probabilistically. By sampling this probabilistic model, it is possible to generate new texts as well as to control various parameters related to the characteristics of the generated text.

Finally, once the text has been generated, it is necessary to define how it will be mathematically represented and which classifiers will be used to discriminate between true and false reviews, and thus test the effectiveness of GPT2 as a generator of fake reviews. These aspects are reviewed in Section 2.3.

2.1. Fake review generation

Fake opinions refer to all misleading opinions shared in a digital environment that do not reflect the genuine opinion of their author (Hunt, 2015). Some authors put the presence of these opinions in certain areas as high as 33 % (Salehi-Esfahani and Ozturk, 2018), which poses a risk to consumer trust and limits the development of e-commerce. Many of the fake reviews come from online sellers themselves, who post positive reviews of their products and negative reviews of competitors, or from review exchange platforms.

Many cases have come to light in recent years, such as TripAdvisor in 2012, denounced by the UK Advertising Standards Authority for including unverified reviews; Samsung in 2013, condemned by the Taiwan Federal Trade Commission for posting false negative reviews against its competitor HTC; or Mafengwo.com in 2018, also for posting false reviews against competitors. However, many other cases still remain hidden. Currently, fraudulent reviews continue to grow rapidly, with a widespread presence in many areas of e-commerce. As a result, the identification of fraudulent reviews has become an emerging field of research (Wu et al., 2020).

One of the most commonly used methods for the proliferation of online fake reviews on the Web is crowd-turfing (Rinta-Kahila and Soliman, 2017), in which companies pay in exchange for obtaining positive ratings written by workers. However, this method is not the most appropriate for large-scale attack systems, as it comes at a financial cost. An automated fake review generation system using deep learning techniques is a much more efficient method of performing massive attacks. Deep learning algorithms have provided huge advances in the field of natural language processing, so that it can be said that there is a competition between generators, always improving their capacity to resemble truthful reviews, and discriminators, also improving their capacity to detect fake reviews (Adelani et al., 2020; Köbis and Mossink, 2021). Generally, deep learning classifiers using CNN layers, LSTM layers, or a hybrid combination have achieved good accuracy results when applied to annotated datasets with human-generated fake reviews. However, the ability of automatic generators to resemble truthful reviews has improved so much that they represent a real threat to current discriminators. In the next sections, we detail different existing methods both for generators and classifiers.

2.2. Review of text generators

The number of deep learning approaches for generative models has increased in recent years. Recurrent Neural Networks (RNN) were among the earliest generative models (Rumelhart et al., 2013). These networks include an internal memory (hidden states) so that they can recall previous outputs, which make them ideal for modelling sequences. However, there is an issue with this design when modelling long-term dependencies: the exploding/vanishing gradients. The problem of exploding/vanishing gradients is caused by the backpropagation algorithm that neural networks use during training. As the backpropagation algorithm moves from the output layer to the input layer, the gradients tend to become either smaller and smaller and closer to zero or larger and larger, leading to larger and larger weights. As a result, the gradient descent never converges to the optimum.

Long Short-Term Memory (LSTM) networks, a RNN model variation, are composed of memory cells that can discover long-term dependencies, so they can learn to preserve only relevant information along sequences and avoid the problem of the exploding/vanishing gradients (Chen et al., 2016; Chung and Sohn, 2020). Several studies have shown how RNN and LSTM networks have been widely used for content generation in recent years (Pawade et al., 2018; Goodfellow et al., 2020).

There are basically three main approaches for text generation: Generative Adversarial Networks (GANs), Variational AutoEncoders (VAE), and Transformer-based generative models (Selvarajah and Nawarathna, 2021). Fig. 1 illustrates a basic block diagram of the three approaches.

GANs are made up of a generative component and a discriminative component that are trained simultaneously. Fig. 1(a) shows its basic scheme, consisting of a generator and a discriminator, both based on neural networks, competing against each other in a zero-sum game. The network converges to the Nash equilibrium point, which is the point at which neither the encoder nor the decoder has a chance to improve. Once the network is trained, the decoder is able to generate text from noisy input. This approach has been widely used for image generation; nonetheless, several GAN variations based on GAN sequences, such as SeqGAN or LeakGAN, have been proposed for text generation (Yu et al., 2017; Guo et al., 2018). The main limitation of GAN techniques is that the Nash equilibrium point is difficult to reach, because the encoder and decoder compete non-cooperatively. They also suffer from the

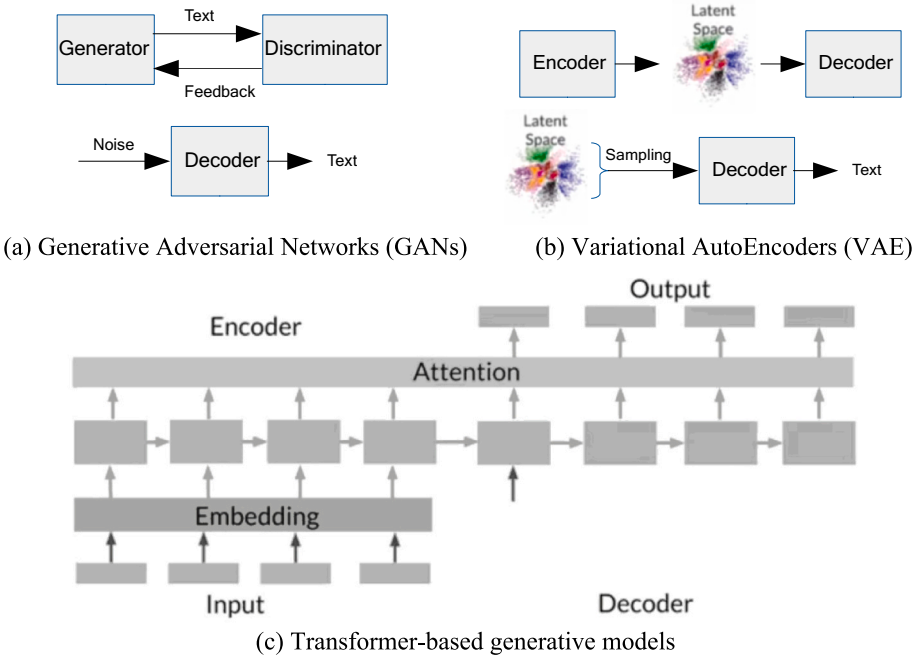

Fig. 1. Block diagram of the main approaches for text generation.

generator-discriminator imbalance in the sense that the discriminator's job is easier than the generator's, so the discriminator tends to achieve very high accuracy at the cost of poor generation by the encoder (Karras et al., 2020).

VAE have grown in popularity since these unsupervised models can operate with unlabeled data and may infer meaningful latent codes from texts. It is made up of an encoder and a decoder that encode input into latent variables with a probability distribution to preserve the integrity and continuity of the latent space (Kingma and Welling, 2014). Fig. 1(b) shows its basic scheme: The encoder generates a probability distribution of the latent variables from which the decoder regenerates the original input. Once the variational autoencoder has been trained, by sampling in the latent space, it is possible to generate a new text using the decoder. Skip-VAE has been proposed as one of the VAE models for text generation (Dieng et al., 2019). The main limitation of these methods lies in the difficulty of accurately generating the probability distribution of the latent space, which compromises the generation of high quality continuous sentences (Kim et al., 2018).

Transformer-based text generation methods have proliferated in recent years, as this novel architecture provides a solution to the problems of gradient leakage when modelling long-term dependencies (Choi et al., 2022). As shown in Fig. 1(c), the transformer architecture incorporates a self-attention layer in both the encoder and the decoder, so that they can look at other words in sequence as it encodes/decodes a specific word. Based on self-attention processes, this architecture eliminates recurrence and convolutions, enabling their training on larger datasets. The most widely reviewed model for text generation based on this architecture is OpenAI's Generative Pre-trained Transformers (GPT2) (Radford et al., 2019). GPT models are designed to pretrain one-directional representations, and they are built using a decoder-only transformer-based architecture. The largest trained model includes 1.5 billion parameters and has outperformed state-of-the-art techniques on natural language processing challenges. As all these models are pre-trained and publicly available, they represent a significant threat in the field of misleading reviews, as anyone can retrain the model on a specific field and use them for malicious purposes. GPT2 has been used extensively for text generation (Das and Verma, 2020) and review generation (Salminen et al., 2022). BERT (Bidirectional Encoding Representations of Transformers) is another transformer-based model that has been frequently utilized for language interpretation tasks, although some recent works have employed this model for text generation (Devlin et al., 2019). The primary distinction between these two models is that BERT uses transformer encoder blocks, while GPT-2 is built using transformer decoder blocks. The main problem with transformer-based models is that the attention mechanism adds more weights to the model and they are difficult to train because they require a very large input corpus (Li et al., 2020). Fortunately, it is possible to download pre-trained versions of these models.

In terms of controlled text generation, a disadvantage of GPT-2 is its lack of high-level semantic control in language generation. That means that this model does not have a parameter by which to begin generating text, so the first word could be any word. The prefix token and the truncation token are parameters that regulate the first and last words of the text, allowing for slight control over the text's generation. GPT2 lacks this option, although a related model, *GPT2-simple*, available for installation from a Python library (https://pypi.org/project/gpt-2-simple/) includes this possibility. A summary of all these techniques for text generation is show below in Table 1.

Table 1
Summary of text generators.

Main architecture	Model	Main characteristics	Ref.
RNN	RNN	Internal memory (vanishing gradient problem)	(Rumelhart et al., 2013)
	LSTM	Can discover long term dependencies	(Chen et al., 2016)
GAN	GAN	Adversarial training	(Goodfellow et al., 2020)
	SeqGAN	GAN sequences	(Yu et al., 2017)
	LeakGAN	Address the problem of long text generation	(Guo et al., 2018)
VAE	VAE	Can operate with unlabelled data	(Kingma and Welling, 2014)
	Skip-VAE	Address the "latent variable collapse" problem	(Dieng et al., 2019)
Transformers	GPT2	Decoder block, one-directional representations	(Radford et al., 2019)
	BERT	Encoder block, bidirectional representations	(Devlin et al., 2019)

2.3. Review of document representations and classifiers

Many authors have studied various methods for detecting fake reviews over the years. Some of the techniques take into account the feelings and emotions depicted in the text (Melleng et al., 2019), or the relationships between reviews, reviewers, and reviewed stores (G. Wang et al., 2018; Z. Wang et al., 2018). However, the majority of the proposed techniques are concerned with extracting informative features from review content, the bag of words approach being the simplest. Basically, the bag of words approach represents documents as a feature vector where each feature is a word and the feature's value if the number of occurrences of each word within the document. The main limitation of the bag-of-words approach is that it provides a sparse representation of documents, i.e., a vector with many zeros, which is incompatible with neural networks, which perform far better with denser representations (Grzeça et al., 2020).

Word-based coding methods solve this problem by capturing syntactic and semantic similarity and relationships between words in an n-dimensional vector. Negative sampling is used in Word2Vec to learn embedding weights by estimating words based on their context (Mikolov et al., 2013). It was one of the first notable word embeddings developed and has been widely employed in natural language processing applications (Yilmaz and Toklu, 2020). Global Vectors for Word Representation, or GloVe, is another major embedding approach (Pennington et al., 2014). Unlike Word2Vec, this methodology generates vectors by combining the local context window technique with global matrix factorization. Finally, FastText enables the model to learn embeddings for OOV (Out Of Vocabulary) words since each word is represented as a bag of n-character frames (sub-words) and is computationally less expensive (Joulin et al., 2017). Word embeddings are widely used in several research fields, including natural language processing tasks such as emotion analysis for text categorization or spam detection (Melleng et al., 2019). However, despite capturing the meaning and context of the word within the sentence, these techniques yield static representations of words.

Unlike traditional word embeddings, neural language models may generate contextual word embeddings, which represent a word in a sentence with a particular vector dependent on its meaning. The Neural Network Language Model (NNLM) is a technique that simultaneously learns a distributed representation for each word as well as the probability function for word sequences (Bengio et al., 2003). Another example of this architecture is ELMo (Peters et al., 2018), which represents not only the dynamic features of word usage, but also how these qualities vary throughout linguistic context.

Other models previously mentioned as text generators can also be used as classifiers. This is the case of BERT and OpenAI GPT2, previously discussed as text generators. Another noteworthy embedding approach is USE (Cer et al., 2018), a model also based on transformer architecture that can build sentence embeddings taking into account both word order and the identification of the remaining words in the sentence. Table 2 summarizes the document representation methods mentioned above.

Word-based representations require a neural network to capture the semantic representation of texts and their sequential sequence. The aim of the neural network is to learn phrase and bigger text representations from dispersed word representations. Therefore, the majority of them are based on sequential models such as LSTMs or a mixture of CNNs and RNNs (Table 3). Several more sophisticated neural networks have been developed, including an RCNN employing Word2Vec embedding (Lai et al., 2015), or a bidirectional Gated Recurrent Neural Network (Bi-directional Average GRNN) (Ren and Ji, 2017). Garcia-Silva et al. (2019) propose a CNN for bot identification that has been tested using Word2Vec, GloVe, and FastText. Context-based encoding approaches, unlike word-based representations, do not require a convolutional or recurrent network since they acquire a representation of the complete document by themselves. As a result, a dense neural network (DNN) is sufficient for the classification task (Selvarajah and Nawarathna, 2021).

Table 2
Summary of document representation models.

Main architecture	Model	Main characteristics	Ref.
Word embedding	Word2Vec	Negative sampling	(Mikolov et al., 2013)
	GloVe	Global matrix factorization (global vectors)	(Pennington et al., 2014)
	FastText	Bag of n-character frames, includes OOV tokens	(Joulin et al., 2017)
Neural language models	NNLM	Contextual embeddings	(Bengio et al., 2003)
	ELMO	Complex characteristics of word & how these uses vary across linguistic contexts	(Peters et al., 2018)
Transformer architecture	BERT	Encoder block, bidirectional representations	(Devlin et al., 2019)
	GPT2	Decoder block, one-directional representations	(Radford et al., 2019)
	USE	Word order & identification of the remaining words in the sentence	(Cer et al., 2018)

Table 3
Summary of NN architectures.

Embedding method	NN	Reference
Word-based encoding	RCNN	(Lai et al., 2015)
	Bi-directional average GRNN	(Ren and Ji, 2017)
	CNN	(Garcia-Silva et al., 2019)
Context-based encoding	DNN	(Selvarajah and Nawarathna, 2021)

It should be noted that most of the previous classifiers were trained with manually-generated deceptive datasets, but they have not been tested with bots.

Although the above methods of document representation and classification have been used in many studies related to natural language processing, no study to date has used them to characterize text generators as false review generators. Classification work in this context has mainly focused on discriminating manually-generated fake reviews from real reviews. However, given the rise of new automatic generation tools, it is of vital importance to characterize the performance of automatic text generators and their implications for existing classification methods.

3. Research framework

3.1. Research design and hypotheses

Previous studies have shown how the proliferation of fake reviews has been boosted by the development of automated text generators (Köbis and Mossink, 2021). Some of the most sophisticated models, such as GPT-2 or BERT, have achieved state-of-the-art performance on a wide range of tasks, including natural language understanding (NLU), sentence classification, named entity recognition, and question answering, transforming the natural language processing landscape. One property of these language models is the possibility of fine-tuning, i.e., retraining the models on a smaller corpus of text on a particular topic. In this way, text generators learn to generate more specialized opinions on that topic (Salminen et al., 2022). This ability of text generators to generate grammatically correct opinions coupled with their ability to specialize on certain topics may hinder the classification task.

Another aspect to consider is that most false opinion classifiers have been trained on manually-generated fake opinions, which are those that are readily available in public databases. However, the logic used by automatic generators is not the same as that used by individuals writing

false opinions, so it is to be expected that the classifier will behave worse to a style of reviews that has not been seen before. Hence, previous considerations lead to the following hypothesis:

H1. Automatically-generated deceptive reviews worsen the performance of deceptive classifiers more than manually-generated deceptive reviews.

These language models are pre-trained on vast volumes of raw or unlabeled textual information in order to create language models that can be quickly deployed to natural language-based applications with little or no fine-tuning (See et al., 2019). By using fine-tuning, the text created by these language models resembles its training data, making it nearly hard for the human eye to distinguish between genuine and false statements.

When using a text generator to generate fake reviews, there are two possibilities: either generate them from real user reviews or generate them from fake reviews, also generated by users. From the perspective of a manipulating agent, it makes more sense to generate them from fake reviews that already impinge on negative characteristics of competing products. However, given the ability of text generators to mimic the body of text being fine-tuned, it is expected that false opinion classifiers will perform better when the generator is trained on false opinions than on true opinions. Hence, previous considerations lead to the following hypothesis:

H2. The performance of the classifier is better when automatically-generated deceptive reviews are generated from original deceptive reviews than from original truthful reviews.

As mentioned above, the *GPT2-simple* model that we will use for this study also has certain control parameters. Among them, temperature is a parameter that measures the randomness or inventiveness of the generated text in the range [0–1]. As the temperature drops, the generated text becomes predictable and repetitive, with fewer random completions; when it approaches zero, the model becomes deterministic and repetitive (Solaiman et al., 2019). Furthermore, prior studies have also shown too that a higher value for this hyperparameter ($t > 0.5$) yields better outcomes, i.e., deceptive reviews more similar to the real ones (Das and Verma, 2020). In addition, an increase in the innovation of the review will result in an increment in the diversity of the reviews and their vocabulary, generating reviews of a wider scope and therefore worsening the performance of the classifier. Hence, previous considerations lead to the following hypothesis:

H3. Increasing the inventiveness of automatically-generated reviews worsens the performance of the classifier.

The credibility of a review is conditioned by several factors such as the length of the review, the use of certain adjectives, or the emotional bias of the point of view. Some authors present a study of review credibility as a function of the sentiment of the review, among other parameters (You et al., 2020). The study developed by Chung and Zeng (2020) also demonstrates how online opinions influence users in one way or another depending on the emotion expressed in the opinion. For example, it shows that fear and anger emotions have a greater influence on users. A second study, developed by Banerjee and Chua (2021), demonstrates that other parameters such as specificity or exaggeration are positively and negatively related, respectively, to the degree of authenticity with which users perceive the review. Other studies also show how the emotion of the review affects how these reviews influence users, concluding that emotions such as fear and anger (associated with negative reviews) tend to have more influence on users, appearing more truthful Chung and Zeng (2020). Therefore, the sentiment expressed in the review has an important influence on whether that review is identified as truthful or deceptive (Du et al., 2020). Hence, previous considerations lead to the following hypothesis:

H4. Negative sentiment reviews tend to be more credible than positive sentiment reviews, which worsens the performance of the classifiers.

3.2. Methodology

This study examines how indistinguishable automatic and manually-generated fake reviews are using a text-generator such as GPT2 and two classifiers of misleading reviews. The first classifier will be built using a word-encoding representation followed by a Bi-LSTM neural network, similar to the one proposed by Graves and Schmidhuber (2005). The word-embedding method used will be FastText, as this method also takes into account OOV tokens, unlike the other word embeddings. Unlike the previous one, the second classifier will be built using a context-based encoding method followed by a DNN (Selvarajah and Nawarathna, 2021). The context-based encoding method used will be the Universal Sentence Encoder, or USE, which is a model based on a transformer architecture that consider both the word order and the identification of the remaining words in the sentence.

Three different datasets will be used to carry out the experiments, comparing the classifier results for both manually made and automatically-generated false reviews. As shown in Fig. 2, true reviews are always those captured in the three starting datasets, while fake reviews can come either from the datasets used or from those automatically-generated by GPT-2.

The *GPT2-simple* model used includes the possibility of generating control tokens to guide text generation and control hyperparameters such as temperature.

The temperature hyperparameter measures the degree of inventiveness or randomness of the text. This parameter can be modified in the range [0–1] and the higher the value, the more varied the reviews will be from each other.

The control tokens indicate whether the generated review is positive or negative, and the end of the generated text. The initial tokens "⟨positive⟩" and "⟨negative⟩" are added at the beginning of the review depending on their sentiment. The sentiment of each review is known in advance because the selected databases include this information as part of the metadata. At the end of each review, the token ⟨endoftext⟩ is added to indicate that this is the end of the review. After finetuning, this model will be able to generate the desired number of fake reviews with a controlled sentiment polarity. The combination of pre-training and supervised finetuning has become the best practice for state-of-the-art results (Radford et al., 2019).

4. Datasets and experiments setup

4.1. Datasets

The experiments will be carried out using three datasets from three different fields: restaurants, home appliances, and hotels, as indicated in Table 4.

D1 refers to the "Deceptive Opinion Dataset" and was obtained through Kaggle (https://www.kaggle.com/rtatman/deceptive-opinion-spam-corpus, https://myleott.com/op-spam.html) and consists of 1600 reviews from TripAdvisor and Mechanical Turk from 20 different hotels in Chicago; it is divided into positive and negative evaluations, with the two classes balanced. D2, the home dataset, contains 2100 reviews on household products that have been collected from three separate departments of the "Amazon reviews dataset": Home, Home Entertainment, and Home Improvement. It is also available on Kaggle (https://www.kaggle.com/lievgarcia/amazon-reviews). Finally, the YelpChi Restaurant dataset (http://odds.cs.stonybrook.edu/yelpchi-dataset/) includes 31,000 reviews of Chicago restaurants. This dataset contains ratings, a date, and product and user information. Only 4000 reviews were chosen at random to maintain the dataset the same size as the others. All three datasets report on the polarity of the reviews.

Fig. 3 shows some examples of truthful, deceptive and GPT2-generated deceptive reviews from the dataset D1 (Chicago hotels), one

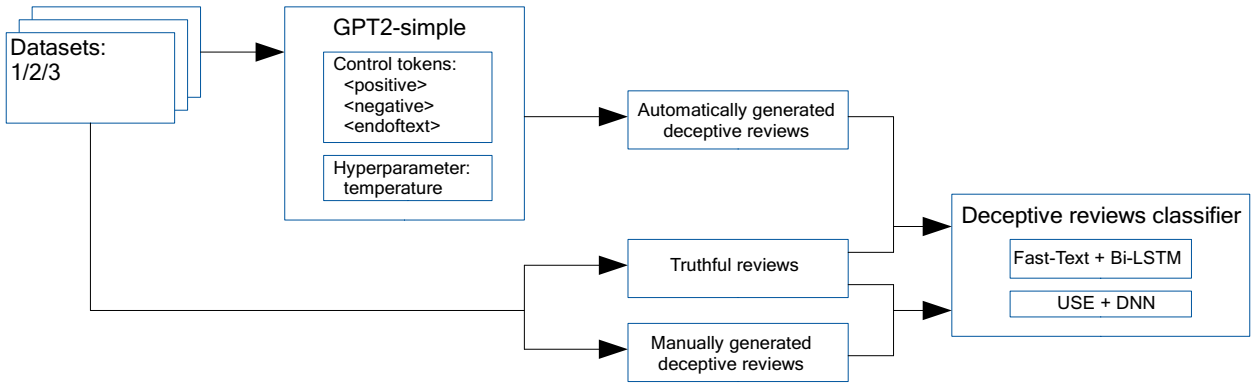

Fig. 2. General scheme.

Table 4
Used datasets.

Dataset name	Content	N° reviews	Source
D1	Chicago hotels	1600	TripAdvisor, MTurk
D2	Home appliances	2100	Amazon
D3	Chicago restaurants	4000	Yelp

positive and one negative. The last two rows correspond to the fake reviews generated by GPT2. It should be noted that even though they are generated automatically, they are completely legible and understandable, preserving the logic of the sentence. Both reviews have been generated using a maximum review size of 800 words, with the temperature hyperparameter set at 0.7 and each with a different sentiment token (⟨positive⟩, ⟨negative⟩).

4.2. Experiment setup

In this section, the setup experiments carried out for each of the four research hypotheses mentioned above will be defined.

4.2.1. Experiment 1

The aim of this first experiment is to test the first hypothesis, H1, i.e., whether automatically-generated deceptive reviews worsen the performance of classifiers with respect to manually-generated deceptive reviews. The initial stage will be to split each dataset into training and test datasets: "*DX_train*" and "*DX_test*" (80 % and 20 %, respectively, with $X = 0, 1, 2$) and train the classifier. Once the classifier is trained, it will be tested with two different test datasets, the original manually-generated dataset "*D_test*", and the automatically-generated dataset "*D_test_GPT2*". The latter dataset contains the original true test reviews, but replaces the original fake reviews with fake reviews automatically-generated by GPT2 using the original true reviews.

For both test data sets, the accuracy of each classifier will be compared. Note that, in terms of the number of true/misleading and positive/negative reviews, each data set is appropriately balanced. This procedure is shown in Fig. 4 and will be carried out for each of the three selected data sets.

4.2.2. Experiment 2

The objective of this second experiment is to find out how the classifier performance is affected by the fact that the misleading reviews generated by GPT2 were generated either from the misleading reviews of the original dataset or from the truthful reviews. For this purpose, two analogous sub-experiments have been carried out that differ only in how the fake reviews have been generated, as shown in Fig. 5. In case 2A, GPT2 has been fine-tuned with the fake reviews, so that the proposed fake review classifiers are trained and tested using the original true reviews and the fake reviews created by GPT2 using the original fake reviews. In case 2B, GPT2 is fine-tuned with the true reviews, and the classifiers are trained and tested with the fake reviews created by GPT2 from the original true reviews. Thus, the only thing that changes in both cases is the provenance of the reviews against which GPT2 is fine-tuned. Comparison of the results obtained will allow us to test hypothesis H2.

4.2.3. Experiment 3

The third experiment focuses on analyzing how a variation in the inventiveness of the reviews affects the performance of the classifiers. In this case, for each dataset, GPT2 will generate deceptive reviews by varying its temperature control parameter in the range of 0 to 1, in 0.1 intervals, until it obtains 10 sets of automatically generated fake reviews. As in previous experiments, the different generated datasets will be fully balanced and the performance of both classifiers will be tested for the three datasets for each temperature value.

4.2.4. Experiment 4: influence of sentiment on accuracy

The last experiment focuses on demonstrating whether the sentiment of the review affects the perception of truthfulness or deception by online users, and whether reviews with negative sentiment are more credible than those with positive sentiment. In this case, the possibility of controlling the polarity of the opinion generated by the *GPT2-simple* model using the initial token ⟨positive⟩ or ⟨negative⟩ will be exploited, generating as a result a dataset of true and false positive opinions and another dataset of true and false negative opinions. Again, the classification task will be performed with both classifiers and the two datasets of positive and negative reviews, comparing the results obtained for each of them.

5. Results

The results are presented in four tables, one for each of the experiments. All these results have been carried out using the Google Colab platform, running on the NVIDIA Tesla T4 GPU, equipped with approximately 15 GB of available memory. Tables 5, 6, 7 and 8 detail the results obtained for the four proposed experiments with two different classifier models and three datasets. More specifically, "D1", "D2", and "D3" refer to the "Hotels", "Home", and "Restaurants".

Table 5 reveals that, in the experiment associated with the first hypothesis, the automatically generated reviews do worsen the performance of the classifier, i.e., hypothesis H1 is confirmed.

As an example, for dataset 1, the classification of manually-generated fake reviews obtained a result of 0.856 accuracy, while for the GPT2-generated fake reviews classification task this performance gets much worse, transforming the classifiers into almost a random classifier with results of 0.535 for USE and 0.501 for FastText. Similar results are observed in Table 5 for the other two datasets.

Table 6 shows the results of Experiment 2. The first part of the table shows the performance of the text classifiers when GPT2 generates

Dataset truthfulness	Sentiment	Review content
Truthful (Original dataset)	Positive	*"This hotel was very beautiful. I went with a group of 4 friends, and we had a blast. It was very centrally located and was close to all sorts of transportation. The Continental breakfast was magnificent! I didn't see much of the staff but all of the men at the front door were extremely helpful."*
	Negative	*"I stayed here with a dozen of my friends, and more than half of us got eaten alive by bed bugs! We stayed there from September 28 - September 30th, but the symptoms (i.e. the crazy itchiness) didn't come until Sunday night when we all got home. I complained to the hotel and basically all they're gonna do is cover the cost of our dry-cleaning bills. Honestly, they should reimburse us because it was not worth it at all. DO NOT STAY AT THIS HOTEL!!!"*
Deceptive (Original dataset)	Positive	*"Fairmont Chicago was a great choice for my wife and I. We couldn't resist the elegant Aria Restaurant and the accommodations. The room was impeccably clean, and the city view was utterly amazing. From the moment we checked in, the staff produced a positive atmosphere with genuine hospitality. My wife even suggested we should have had our wedding reception in their beautiful ball room. Overall, I would rate Fairmont Chicago a 10/10! "*
	Negative	*"Recently stayed at the James Chicago and will NOT be returning anytime soon. The staff is unfriendly to say the least and the prices are outrageous. Also, not a fan of the fact they allow pets as guests. There wasn't enough space in the gym and the lack of variety in equipment limits your options severely. Ultimately I was very unimpressed with the services as a whole and would recommend you avoid this establishment."*
Deceptive (GPT2-generated)	Positive	*"We stayed at the Allegro in July for 4 nights for a business trip. The room was spotlessly clean, and the bed was great. I couldn't wait to get out for dinner. The location was great. Two blocks from the Mag Mile and a block from the Washington El train. The front desk and front desk staff were very friendly and efficient. We had coffee and juice in the lobby and the bar and restaurant both have water and juice. The breakfast buffet was very good, but the dinner buffet was a bit overwhelming. For a 4-star hotel, it should be a good experience."*
	Negative	*"Well, our check-in was great, but then we got into our room. The carpet was drenched and stained near the bathroom because the shower leaks water onto the floor since there are no doors on it. The mini bar was broken and put an extremely foul scent into the room. Our final hotel bill reflected 2 charges that were incorrect....1 for the mini bar which we did not utilize, and the second was an up charge for a bottle of wine ($69 instead of $50). When my husband went to the front desk to have it corrected, he was met with a very unfriendly and unapologetic employee. It was adjusted, but not without confrontation. Overall, the location of the Hard Rock was nice but it will be the last time that we stay in one,"*

Fig. 3. Corpus samples.

deceptive reviews from the deceptive original reviews (Experiment 2A), while in the second part of the table, GPT2 generates deceptive reviews from the truthful original reviews (Experiment 2B).

A value of 0.881 can be observed for the dataset D1 and the FastText classifier (Experiment 2A), versus a value of 0.649 with the same classifier and a different dataset when these deceptive reviews are generated from the truthful reviews (Experiment 2B). This comparison reveals that the GPT2-generated text tends to emulate the text it was fine-tuned with, which complicates the classification task quite a bit when the GPT2 deceptive reviews are quite similar to the original truthful reviews (Experiment 2B).

Table 7 shows the results of the classification models against reviews with different degrees of inventiveness (generated at different temperature points). It can be seen that, as the temperature increases, the accuracy of the classifier decreases, as the text becomes more diverse, making the classification task more difficult and thus proving the confirmation of hypothesis H3: "*Increasing the inventiveness of automatically-generated reviews worsens the performance of the classifier*."

Finally, Table 8 shows the results of the last experiment, Experiment 4, in which positive and negative text classifications were conducted separately to demonstrate whether sentiment affects the performance of classifiers. Findings reveal that the positive sentiment of the review slightly hinders the classification task, obtaining slightly lower accuracy results than those with a negative sentiment (e.g. 0.718 vs. 0.731 for the case of dataset D1 and FastText classifier model). However, the difference in classifier performance is so small between the positive and negative reviews classification that hypothesis H4 is not supported.

It is worth noting that the best classification results are always found in the D3 dataset since it is the largest. In contrast, the worst ones are for dataset D2, since, unlike the other two datasets, this one has a more diverse scope of reviews, making the classification task harder. This wider scope is due to the fact that, as mentioned above, this dataset was created by bringing together 3 different categories of reviews (Home, Home Entertainment and Home Improvement), all of them related to the main topic home but with a more diverse vocabulary. Overall, the results are quite similar for both classifiers, with FastText's results slightly better.

Figs. 6, 7, 8 and 9 show the evolution per epoch of the test accuracy during the training stage for each one of the four experiments detailed before. Fig. 7 shows that for Experiment 2 and dataset 3, both the

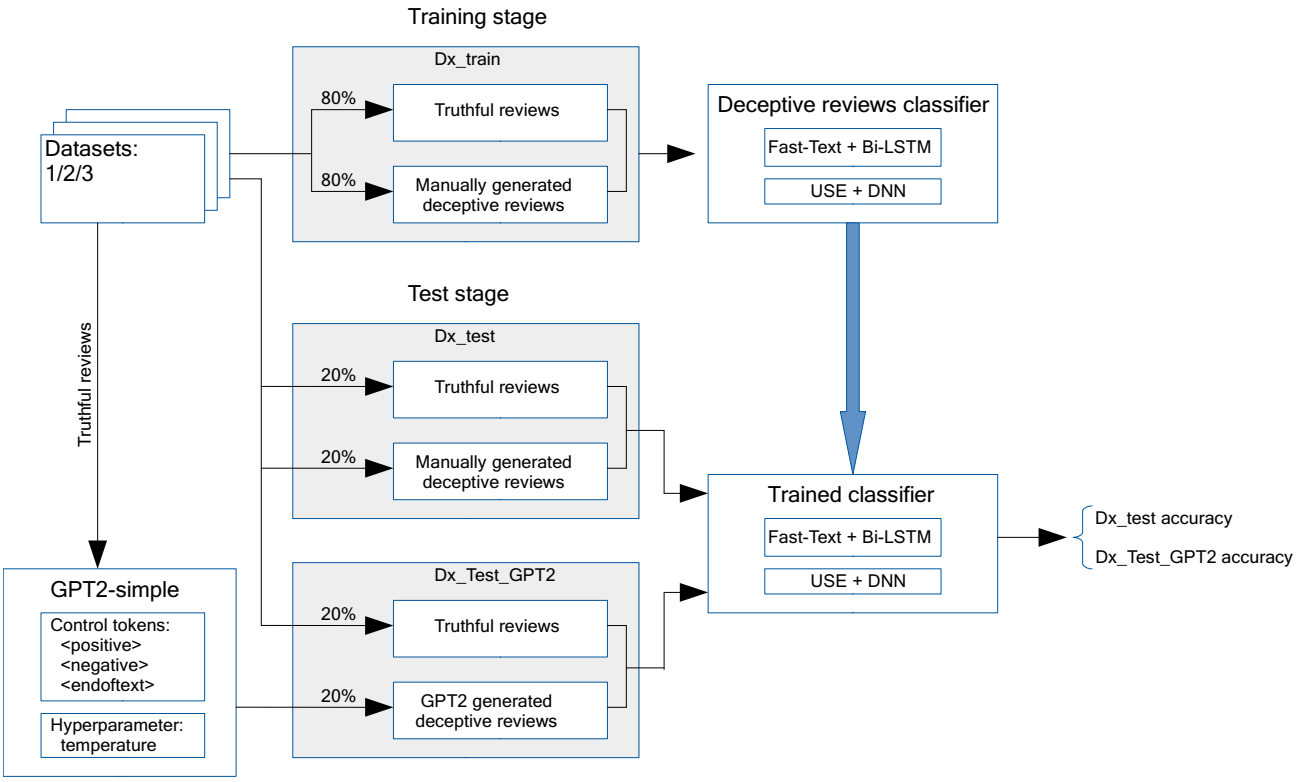

Fig. 4. Experiment 1 setup.

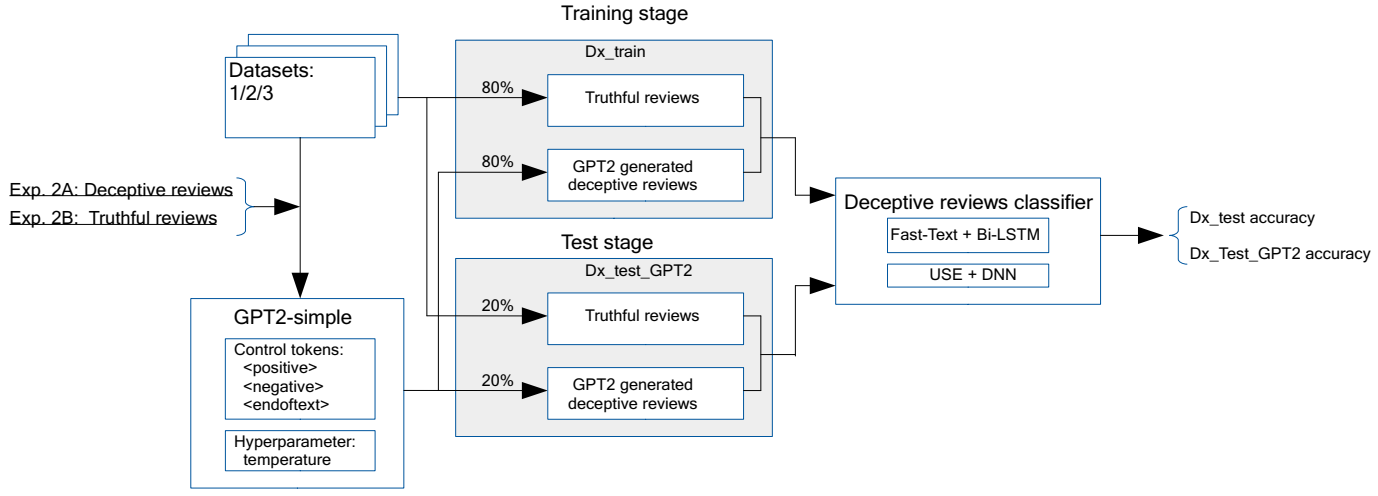

Fig. 5. Experiment 2-setup.

Table 5
Experiment 1 results: manually VS Automatic generated deceptive reviews.

Experiment 1		D1	D2	D3
Manually-generated deceptive reviews	FastText	0.856	0.697	0.735
	USE	0.856	0.679	0.751
GPT2-generated deceptive reviews	FastText	0.501	0.495	0.537
	USE	0.535	0.487	0.541

Table 6
Experiment 2 results: deceptive reviews generated from truthful VS deceptive reviews.

Experiment 2		D1	D2	D3
Generated from deceptive (2A)	FastText	0.881	0.775	0.877
	USE	0.872	0.796	0.862
Generated from truthful (2B)	FastText	0.649	0.654	0.863
	USE	0.649	0.671	0.810

reviews generated from deceptive reviews and the truthful reviews obtain great classification results, and this may be due to the size of the dataset. Being a large dataset, training on a large amount of text can ease the classification task, obtaining good classification results even when the fake reviews are generated to resemble the real ones. In terms of convergence, we also observe a more linear convergence for the USE classifier than for FastText.

The findings yield four important conclusions: First, automatically-generated fake reviews are more difficult to detect than human-

Table 7
Experiment 3 results: effect of inventiveness.

Experiment 3: temperature		0.1	0.3	0.5	0.7	0.9
D1	FastText	0.975	0.856	0.760	0.691	0.643
	USE	0.890	0.757	0.745	0.697	0.655
D2	FastText	0.971	0.895	0.870	0.850	0.725
	USE	0.983	0.970	0.895	0.783	0.716
D3	FastText	0.993	0.980	0.954	0.864	0.690
	USE	0.986	0.961	0.923	0.829	0.701

Table 8
Experiment 4 results: effect of sentiment.

Experiment 4: polarity		D1	D2	D3
Positive	FastText	0.718	0.666	0.854
	USE	0.668	0.666	0.820
Negative	FastText	0.731	0.716	0.879
	USE	0.681	0.691	0.853

generated ones; second, GPT2 generates text very similar to the text it is fine-tuned with; and third, the degree of inventiveness of the reviews affects the variety of the reviews and thus the classification performance, while the sentiment of the review does not influence to a large extent. It is concluded that these results support Hypotheses H1, H2, and H3; while hypothesis H4 is discarded.

6. Discussion and implications

The identification and analysis of deceptive reviews on the Internet have attracted the attention of the research community over the years (Yao et al., 2017). The proliferation of fake reviews has grown with the development of new natural language processing techniques and new transformer-based text generation models. Thus, understanding how language generation models work and their internal characteristics and parameters is critical to know what threats will need to be addressed.

This paper analyzes the different parameters of language generation models and the performance of one of the most cited text generation models: GPT2. In addition, it raises and answers four hypotheses of great current research interest, performing real experiments with two text classifiers previously proposed in the literature on false review identification. Specifically, we study whether parameters such as the degree of inventiveness or the sentiment of the generated deceptive reviews affect the veracity with which the reviews are perceived by a text classifier; as for the performance of GPT2, we study how it tends to mimic the training text and how these automatically-generated reviews present an even greater threat, as they are more difficult to identify as fake than manually-generated ones.

6.1. Theoretical implications

Previous studies have tested the effectiveness of GPT2 as a text generator, but it has not been studied in the specific domain of fake reviews (Radford et al., 2019). Based on four quantitative analyses, this study contributes to academic research with a new approach to data analysis, providing new insights into the performance of the GPT2 text generator in relation to different hyperparameters. The theoretical contributions of this article are fourfold.

In the first analysis, the ease with which GPT2-generated text can be passed off as truthful reviews has been studied. The results shown in Fig. 5 are very conclusive, with a detection phase accuracy of 0.5 for misleading reviews generated by GPT2 versus 0.85 for deceptive reviews generated manually in dataset 1, and with very similar results for the other datasets. These results demonstrate the threat posed by text generators when used for malicious purposes (Salminen et al., 2022) and support hypothesis H1.

In the second analysis, one of the most significant behaviors of the generator, which is its ability to resemble the input text, has been demonstrated through two experiments (See et al., 2019). In this case, the difference in accuracy when generating false automatic reviews from true manual reviews or from false manual reviews has been obtained. The result obtained supports hypothesis H2 that GPT2 tricks the fake review detectors much more easily if it is tuned from true reviews. The drawback of using GPT2 in this way is that the generated reviews would not be used to defame competitors' products, although they could be

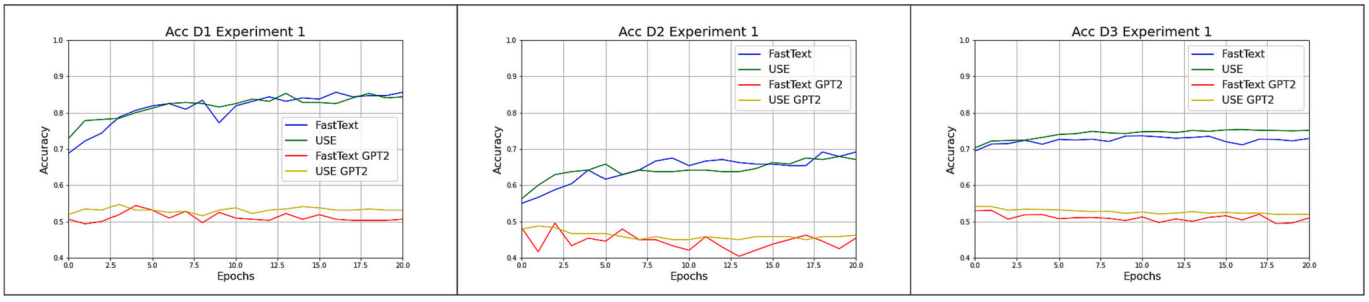

Fig. 6. Experiment 1: results.

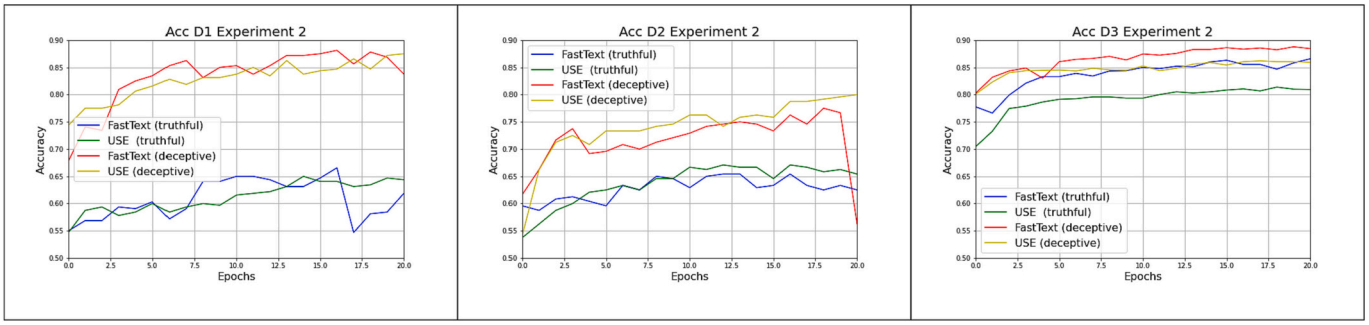

Fig. 7. Experiment 2: results.

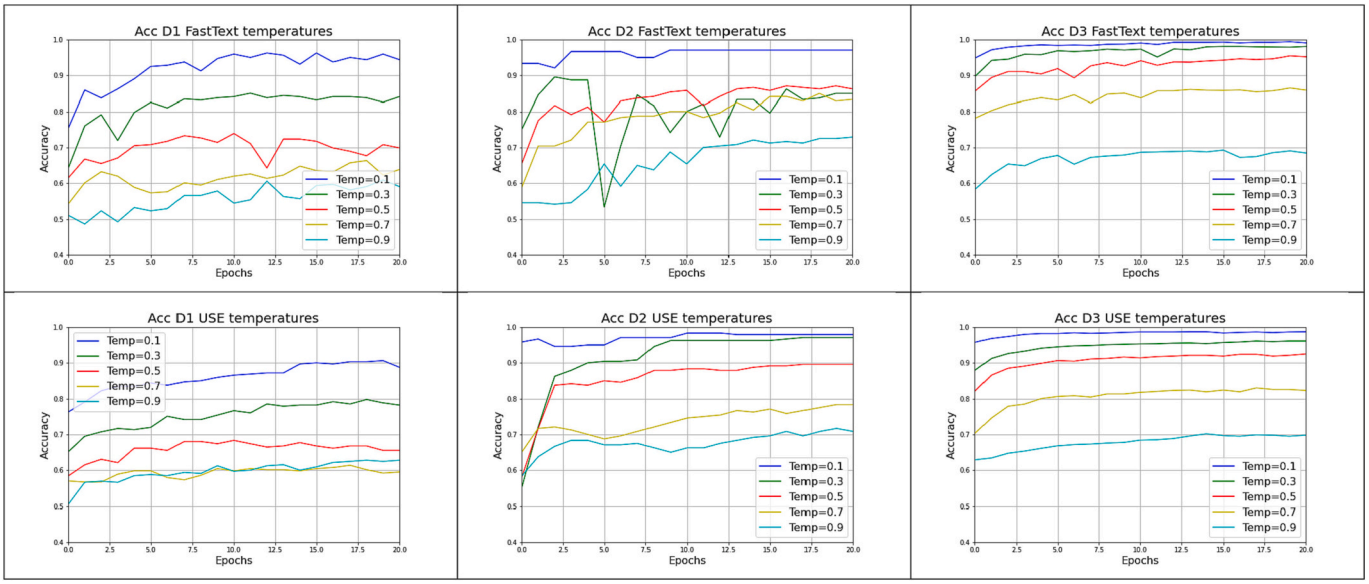

Fig. 8. Experiment 3: results.

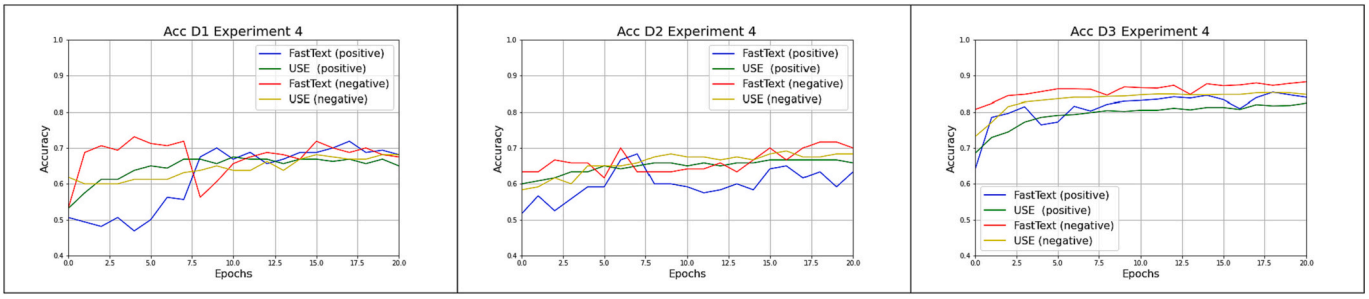

Fig. 9. Experiment 4: results.

used to artificially promote one's own products.

In the third analysis, it is qualitatively demonstrated, firstly, how the variation of one of the internal parameters of the model can significantly change the degree of inventiveness of the generated texts (Solaiman et al., 2019), and secondly, how the increase of the inventiveness parameter influences the level of detection of fake reviews (Das and Verma, 2020). Thus, it is shown that the more sparse and novel the reviews are, the more easily they will be mistaken for a real text, concluding then that hypothesis H3 is also supported.

Finally, the last analysis shows that the effect of opinion polarity of reviews generated with GPT2 is small. Although positive reviews tend to be slightly more credible, worsening the classifier performance, the results are nevertheless inconclusive, so hypothesis H4 is not supported.

6.2. *Practical and managerial implications*

Pre-trained generators, such as GPT2, can be used to generate text in different domains with only prior knowledge about natural language processing and deep learning, since the code is open. This calls into question some ethical concerns about the use of language models and their possible applications. The ultimate goal of AI is to make machines that are indistinguishable from humans, so language is one of the skills that machines should master for their possible interactions with people. However, this ability can also be exploited for manipulation of humans or customers, as it is possible to force machines to reinforce some preconceived ideas or feelings, or to align with a certain view of the discussion and against the other. The threat at this point is twofold: anyone with some prior knowledge in computer science can train and generate manipulated opinions, and can also do so massively, as generators can spread millions of opinions in a short time with some degree of creativity. Moreover, the field is continuously evolving. Currently, GPT3 and ChatGPT are updated versions of GPT2. They are several times larger than GPT2 and trained with a large number of parameters, but the code is not open, as it has an exclusive license. However, new transformer-based language models that mimic their architecture are now being developed and will be released as open source. Despite this situation, linguistic models also represent an opportunity, in the sense that they can also be used to bypass this thread. A new generation of classifiers can be developed by working with generators such as GPT2 and its open source extensions. In addition, generators can also be used to generate new public datasets for use by the scientific community to overcome this problem. Although some databases of deceptive reviews are publicly available, they are limited in size and tend to be associated with a few specific sectors, such as hotels and restaurants. Generators such as GPT2 can be used to provide new databases on a wide variety of products.

This situation gives rise to a race between operators and opinion platform participants, as both have a vested interest in using the latest technology to gain an advantage over the other. Platform operators, such as Amazon or TripAdvisor, can use AI to improve trust in their platforms and services, while platform participants, such as businesses and individuals, can use AI to artificially promote or demote certain products. Academia can play a crucial role in identifying and combating fake opinions generated by automated bots. Researchers in fields such as computer science, data science, and social science can develop methods for detecting and mitigating the spread of fake opinions on online platforms. This can include techniques for identifying bot-generated

content, and for analyzing the spread of misinformation. Additionally, researchers in fields such as law, policy, and ethics can study the implications of bot-generated content and provide recommendations on how to address it. It is important to notice that is not only academia responsibility, but also the responsibility of platform operators and governments to work together to mitigate the spread of fake opinion and misinformation online.

From a legislative point of view, the EU's attempt to address this situation is called the Artificial Intelligence Act. Although it is still in the approval phase, it could be effective in combating fake reviews by setting standards for the use and development of AI and providing monitoring and enforcement mechanisms to ensure compliance. For example, and in the context of this research, an effective measure could be an obligation for platforms to disclose whether a review has been generated by a bot and clearly indicate this to users. This would represent a major step forward for both users and platforms by improving users' trust in shared content. However, this is a delicate regulation that must maintain a balance between protecting citizens and fostering innovation in Europe.

6.3. Limitations and further research

This study addresses the problem of false review generators by analyzing one of the most important generative models and its characteristics, and also by developing two classification models for the discrimination of generated reviews. However, the following limitations should be noted.

First, this study could be extended to more text generators by comparing their performance and efficiency based on their hyperparameters. A second limitation is the variety and scope of the selected data sets. The research could be extended to other types of product or even different areas, for example, fake news, spam text generation, or even more diverse topics, such as poetry generation. Another possible line of future research is to determine whether detection is influenced by the underlying characteristics of the product being analyzed. In general, products can be classified as search or experience products, so the extent to which each type of product is more prone to being counterfeited could be analyzed. On the other hand, it has been shown in the last experiment of this study that there is a small relationship between sentiment and the veracity with which the review is perceived, but the results are not conclusive. Further analysis and more varied tests in relation to hypothesis 4 are also suggested for future work.

In addition, the study could also be extended using a Siamese neural network (SNN) architecture. These neural networks, known for their ability to calculate the similarity between two documents, could be used to calculate the degree of innovativeness of a text, as another hyperparameter beyond temperature. Finally, finding classifiers capable of fighting against the new generation of linguistic models is undoubtedly one of the greatest challenges ahead.

7. Conclusions

This paper follows a quantitative approach to analyze the characteristics and performance of GPT-2 as a deceptive review generator, evaluating its behavior using two different neural network-based classifiers proposed in previous work. The results reveal the threat posed by artificial text generators, as they clearly worsen the performance of the classifiers with respect to manually-generated fake reviews. Furthermore, text generators can be controlled by selecting the polarity of reviews or even the degree of innovation, so websites can be easily flooded with malicious and controlled reviews to mislead consumers. This study demonstrates the ability of GPT-2 to resemble the input text used to fine tune the generator and how the degree of innovativeness worsens the performance of the classifiers. In contrast, the effect of review polarity on classification accuracy is inconclusive.

CRediT authorship contribution statement

A. Perez-Castro: Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Visualization, Writing – original draft. **M.R. Martínez-Torres:** Conceptualization, Methodology, Investigation, Writing – review & editing, Project administration, Funding acquisition. **S.L. Toral:** Conceptualization, Methodology, Formal analysis, Investigation, Resources, Visualization, Writing – review & editing, Supervision.

Declaration of competing interest

None.

Data availability

The data are publicly available and the link to access them is provided in the paper

Acknowledgements

This work was supported by the project Aplicación de Redes Generativas Antagónicas para Combatir la Manipulación de Clientes Online (REACT) Ref. PID2020-114527RB-I00 funded by MCIN/AEI/10.13039/501100011033, by the project "Identificación de opiniones fraudulentas en portales de opinión mediante técnicas de Machine y Deep Learning" US-1255461 funded by Junta de Andalucía, and by the project "Identificación de los Atributos Únicos de los Destinos Turísticos Andaluces desde la perspectiva de los Social Media mediante el uso de técnicas de Text Mining (TURIMEDIA)" PY20_00639 funded by Junta de Andalucía.

References

Adelani, D.I., Mai, H., Fang, F., Nguyen, H.H., Yamagishi, J., Echizen, I., 2020. Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection. April. In: International Conference on Advanced Information Networking and Applications. Springer, Cham, pp. 1341–1354. https://doi.org/10.1007/978-3-030-44041-1_114.

Ahmad, T., Aliaga Lazarte, E.A., Mirjalili, S., 2022. A systematic literature review on fake news in the COVID-19 pandemic: can AI propose a solution? Appl. Sci. 12 (24), 12727.

Banerjee, S., Chua, A.Y., 2021. Calling out fake online reviews through robust epistemic belief. Inf. Manag. 58 (3), 103445.

Bengio, Y., Ducharme, R., Vincent, P., Janvin, C., 2003. A neural probabilistic language model. J. Mach. Learn. Res. 3, 1137–1155. https://doi.org/10.1162/153244303322533223.

Cer, D., Yang, Y., Kong, S.Y., Hua, N., Limtiaco, N., John, R.S., Guajardo-Céspedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., Kurzweil, R., 2018. Universal sentence encoder. available at. https://arxiv.org/abs/1803.11175.

Chen, T., Xu, R., He, Y., Xia, Y., Wang, X., 2016. Learning user and product distributed representations using a sequence model for sentiment analysis. IEEE Comput. Intell. Mag. 11 (3), 34–44. https://doi.org/10.1109/MCI.2016.2572539.

Choi, S., Lee, H., Park, E., Choi, S., 2022. Deep learning for patent landscaping using transformer and graph embedding. Technol. Forecast. Soc. Chang. 175, 121413.

Chung, P., Sohn, S.Y., 2020. Early detection of valuable patents using a deep learning model: case of semiconductor industry. Technol. Forecast. Soc. Chang. 158, 120146.

Chung, W., Zeng, D., 2020. Dissecting emotion and user influence in social media communities: an interaction modeling approach. Inf. Manag. 57 (1), 103108 https://doi.org/10.1016/j.im.2018.09.008.

Das, A., Verma, R.M., 2020. Can machines tell stories? A comparative study of deep neural language models and metrics. IEEE Access 8, 181258–181292. https://doi.org/10.1109/ACCESS.2020.3023421.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North, pp. 4171–4186. https://doi.org/10.18653/V1/N19-1423.

Dieng, A.B., Kim, Y., Rush, A.M., Blei, D.M., 2019. April avoiding latent variable collapse with generative skip models. In: The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, pp. 2397–2405.

Du, X., Zhu, R., Zhao, F., Zhao, F., Han, P., Zhu, Z., 2020. A deceptive detection model based on topic, sentiment, and sentence structure information. Appl. Intell. 50 (11), 3868–3881.

Filieri, R., Alguezaui, S., McLeay, F., 2015. Why do travelers trust TripAdvisor? Antecedents of trust towards consumer-generated media and its influence on recommendation adoption and word of mouth. Tour. Manag. 51, 174–185. https://doi.org/10.1016/j.tourman.2015.05.007.

Garcia-Silva, A., Berrio, C., Gómez-Pérez, J.M., 2019. An empirical study on pre-trained embeddings and language models for bot detection. In: Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), pp. 148–155. https://doi.org/10.18653/v1/w19-431.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Commun. ACM 63 (11), 139–144. https://doi.org/10.1145/3422622.
Graves, A., Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. 18 (5–6), 602–610.
Grzeça, M., Becker, K., Galante, R., 2020. Drink2Vec: improving the classification of alcohol-related tweets using distributional semantics and external contextual enrichment. Inf. Process. Manag. 57 (6), 102369 https://doi.org/10.1016/j.ipm.2020.102369.
Guo, J., Lu, S., Cai, H., Zhang, W., Yu, Y., Wang, J., 2018. Long text generation via adversarial training with leaked information. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32. No. 1.
Hunt, K.M., 2015. Gaming the system: fake online reviews v. consumer law. Comput. Law Secur. Rev. 31 (1), 3–25.
Joulin, A., Grave, E., Bojanowski, P., Mikolov, T., 2017. Bag of tricks for efficient text classification. In: 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference, 2, pp. 427–431. https://doi.org/10.18653/v1/e17-2068.
Karnouskos, S., 2020. Artificial intelligence in digital media: the era of deepfakes. IEEE Trans. Technol. Soc. 1 (3), 138–147.
Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T., 2020. Training generative adversarial networks with limited data. Adv. Neural Inf. Proces. Syst. 33, 12104–12114.
Kim, Y., Wiseman, S., Miller, A., Sontag, D., Rush, A., 2018. Semi-amortized variational autoencoders. In: International Conference on Machine Learning. PMLR, pp. 2678–2687.
Kingma, D.P., Welling, M., 2014. Auto-encoding variational bayes. available at:. In: 2nd Int. Conf. Learn. Represent. ICLR 2014 - Conf. Track Proc., International Conference on Learning Representations, ICLR, 2014 https://arxiv.org/abs/1312.6114v10.
Köbis, N., Mossink, L.D., 2021. Artificial intelligence versus Maya angelou: experimental evidence that people cannot differentiate AI-generated from human-written poetry. Comput. Hum. Behav. 114, 106553 https://doi.org/10.1016/j.chb.2020.106553.
Lai, S., Xu, L., Liu, K., Zhao, J., 2015. Recurrent convolutional neural networks for text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, 29, p. 1.
Li, Z., Wallace, E., Shen, S., Lin, K., Keutzer, K., Klein, D., Gonzalez, J., 2020. Train big, then compress: rethinking model size for efficient training and inference of transformers. In: International Conference on Machine Learning. PMLR, pp. 5958–5968.
Meel, P., Vishwakarma, D.K., 2020. Fake news, rumor, information pollution in social media and web: a contemporary survey of state-of-the-arts, challenges and opportunities. Expert Syst. Appl. 153, 112986.
Melleng, A., Jurek-Loughrey, A., Deepak, P., 2019. Sentiment and emotion based representations for fake reviews detection. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pp. 750–757. https://doi.org/10.26615/978-954-452-056-4_087.
Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. In: 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings. http://ronan.collobert.com/senna/.
Pawade, D., Sakhapara, A., Jain, M., Jain, N., Gada, K., 2018. Story scrambler-automatic text generation using word level RNN-LSTM. Int. J. Inf. Technol. Comput. Sci. 10 (6), 44–53. https://doi.org/10.5815/ijitcs.2018.06.05.
Pennington, J., Socher, R., Manning, C.D., 2014. GloVe: global vectors for word representation. In: EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. Association for Computational Linguistics (ACL), pp. 1532–1543. https://doi.org/10.3115/v1/d14-1162.
Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations. In: NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1, pp. 2227–2237. https://doi.org/10.18653/v1/n18-1202.
Petrescu, M., O'Leary, K., Goldring, D., Ben Mrad, S., 2018. Incentivized reviews: promising the moon for a few stars. J. Retail. Consum. Serv. 41, 288–295. https://doi.org/10.1016/j.jretconser.2017.04.005.
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language models are unsupervised multitask learners. OpenAI Blog 1 (8), 9.
Ren, Y., Ji, D., 2017. Neural networks for deceptive opinion spam detection: an empirical study. Inf. Sci. 213-224, 385. https://doi.org/10.1016/j.ins.2017.01.015.
Rinta-Kahila, T., Soliman, W., 2017. Understanding crowdturfing: the different ethical logics behind the clandestine industry of deception. available at:. In: ECIS 2017: Proceedings of the 25th European Conference on Information Systems.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 2013. Learning internal representations by error propagation. In: Readings in Cognitive Science: A Perspective From Psychology and Artificial Intelligence. Elsevier Inc., pp. 399–421. https://doi.org/10.1016/B978-1-4832-1446-7.50035-2
Salehi-Esfahani, S., Ozturk, A.B., 2018. Negative reviews: formation, spread, and halt of opportunistic behavior. Int. J. Hosp. Manag. 74, 138–146.
Salminen, J., Kandpal, C., Kamel, A.M., Jung, S.G., Jansen, B.J., 2022. Creating and detecting fake reviews of online products. J. Retail. Consum. Serv. 64, 102771.
See, A., Pappu, A., Saxena, R., Yerukola, A., Manning, C.D., 2019. Do massively pretrained language models make better storytellers?. In: CoNLL 2019 - 23rd Conference on Computational Natural Language Learning, Proceedings of the Conference, pp. 843–861. https://doi.org/10.18653/v1/k19-1079.
Selvarajah, J., Nawarathna, R.D., 2021. In: A Lucrative Model for Identifying Potential Adverse Effects from Biomedical Texts by Augmenting BERT and ELMo. Springer, Singapore, pp. 233–247. https://doi.org/10.1007/978-981-33-4355-9_19.
Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J.W., Kreps, S., Newhouse, A., Blazakis, J., Wang, J., McGuffie, K., McCain, M., 2019. Release strategies and the social impacts of language models. available at: http://arxiv.org/abs/1908.09203.
Wang, G., Li, C., Wang, W., Zhang, Y., Shen, D., Zhang, X., Henao, R., Carin, L., 2018. Joint embedding of words and labels for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL, pp. 2321–2331. https://doi.org/10.18653/v1/p18-1216.
Wang, Z., Gu, S., Xu, X., 2018. GSLDA: LDA-based group spamming detection in product reviews. Appl. Intell. 48 (9), 3094–3107.
Wu, Y., Ngai, E.W., Wu, P., Wu, C., 2020. Fake online reviews: literature review, synthesis, and directions for future research. Decis. Support. Syst. 132, 113280.
Yao, Y., Viswanath, B., Cryan, J., Zheng, H., Zhao, B.Y., 2017. Automated crowdturfing attacks and defenses in online review systems. In: Proceedings of the ACM Conference on Computer and Communications Security, pp. 1143–1158. https://doi.org/10.1145/3133956.3133990.
Yilmaz, S., Toklu, S., 2020. A deep learning analysis on question classification task using Word2vec representations. Neural Comput. & Applic. 1–20. https://doi.org/10.1007/s00521-020-04725-w.
You, X., Lv, X., Zhang, S., Sun, D., Gao, S., 2020. Sentiment analysis of film reviews based on deep learning model collaborated with content credibility filtering. In: International Conference on Collaborative Computing: Networking, Applications and Worksharing. Springer, Cham, pp. 305–319.
Yu, L., Zhang, W., Wang, J., Yu, Y., 2017. Seqgan: sequence generative adversarial nets with policy gradient. In: Proceedings of the AAAI Conference on Artificial Intelligence, 31, p. 1.
Zhou, S., Barnes, L., McCormick, H., Blazquez Cano, M., 2021. Social media influencers' narrative strategies to create eWOM: a theoretical contribution. Int. J. Inf. Manag. 59, 102293 https://doi.org/10.1016/j.ijinfomgt.2020.102293.
Zhu, F., Zhang, X., 2010. Impact of online consumer reviews on sales: the moderating role of product and consumer characteristics. J. Mark. 74 (2), 133–148. https://doi.org/10.1509/jm.74.2.133.

Ms. Amparo Perez-Castro was born in Almería, Spain, in 1997. She received the degree in Electronic, Robotic and Mechatronic Engineering from the University of Seville, Spain, in 2020. She is currently a Data Scientist Researcher for the University of Seville. Her main research interests include Natural Language Processing, Deep Learning and Machine Learning.

Dra. Rocío Martínez-Torres was born in Madrid, Spain, in 1973. She received the degree in Business Administration in 1996 and the Ph. D. degree from the University of Seville, Spain, in 2003. She is currently a Full Professor at Business Administration and Marketing Department, University of Seville. She is author or co-author of 50 papers in major international peer-reviewed journals (JCR/JSCR impact factor). Her research interests include intellectual capital and knowledge management, virtual communities and open innovation.

Dr. Sergio Toral was born in Rabat, Morocco, in 1972. He received the M.S. and Ph.D. degrees in electrical and electronic engineering from the University of Seville, Spain, in 1995 and 1999, respectively. He is currently a Full Professor with the Department of Electronic Engineering, University of Seville. He is an author or co-author of 95 papers in major international peer-reviewed journals (with JCR impact factor) and of over 100 papers in well-established international conferences and workshops. His main research interests include eWOM communities, natural language processing and deep learning.