4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15085

Applying NLP techniques to characterize what makes an online review trustworthy

José Carlos Romero[1], María Olmedilla[2], María Rocío Martínez-Torres[3], Sergio Toral[4]
[1]Department of Computer Architecture, University of Malaga, Spain, [2]SKEMA Business School, France, [3]Facultad de Ciencias Económicas y Empresariales, University of Seville, Sevilla, Spain [4]E. S. Ingenieros, University of Seville, Sevilla, Spain.

Abstract

Users spend a significantly amount of time reading and exchanging reviews online in e-commerce and eWOM communities that help them with their purchase decisions. Source credibility theory is gaining more importance as some online reviews are currently being damaged by those fake reviews that promote an untruthful image not only of the products but also of those online websites. Thus, trustworthiness of online reviews is a key aspect not only for the users that want to make more informed decisions regarding the products, but also for the websites whose credibility might be affected. In this regard, this study proposes a classification system using two Natural Language Processing (NLP) models that can predict trustworthy online reviews (helpful and truthful) applied to the product category "Cell phones & accessories" of Amazon. After using a keyword extractor among those trustworthy online reviews we can characterize their most important features. The results reveal that those features are related to brands, physical and technical features and the UX of the mobile phones.

Keywords: *Source credibility; trustworthiness; helpfulness; online reviews; classifier; Natural Language Processing.*

1. Introduction

The importance of source credibility has been extensively investigated for decades. In this regard, Joshua et al. (1986) investigated the relationship between the constructs of trust and credibility. Their experiments were based on the automobile example from Wart and McGinnies (1980) and the trustworthiness manipulation was based on the Kelly's (1973) ideas. The authors used the level of expertise of the source to show that a more trustworthy source is more credible than a less trustworthy.

Currently, the widespread of the Internet has changed the way messages are exchanged and along with it is the electronic word-of-mouth (eWOM) communication, which has changed the way in which users evaluate the credibility of information (Utz et al., 2012). Thus, most studies have focused on how to cope with source credibility in the context of online reviews (Chakraborty, 2019; Hsieh & Li, 2020). Besides, since the rise of social media, assessing perceptions of source credibility among online reviews has become more important (Metzger et al. 2010). Because online reviews are still playing an important role as information source in the consumer decision making process (Shan, 2016). However, given the decontextualization and anonymity of eWOM the concept of source credibility might seem unclear to consumers (Xie et al. 2011). Likewise, much information exchanged across social media is suspicious or sometimes malicious (Zhang & Ghorbani, 2020). In this respect, recent research works show that AI and ML techniques are playing a vital role in detecting online malicious content (Ahmed et al., 2018; Kaliyar et al., 2020; Ozbay and Alatas, 2020).

Consequently, this paper aims at investigating the relationship between the constructs of trust and credibility applying AI techniques to online reviews. First, we statistically model the online review helpfulness as a component of the construct of source credibility. Then, we make an experiment in which the trustworthiness of the online reviews is characterized by combining two NLP classifiers. One of the NLP classifiers predicts the online helpful reviews, and the other one predicts the truthful online reviews. The common online reviews of both classifiers are called trustworthy reviews. We use a keyword extractor to further characterize the features that make an online review trustworthy.

This paper is structured as follows. Section 2 presents the performed methodology for this work including the followed steps and the data analysis. Section 3 describes the obtained results summarized in three main steps. Finally, Section 5 finishes with the conclusions of the study.

2. Methodology

Figure 1 shows the schematic of the developed methodology.

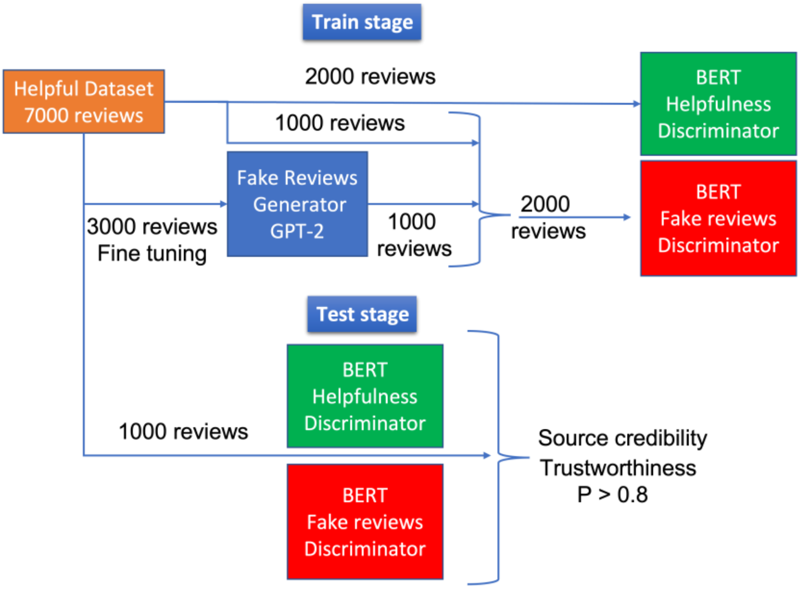
Train stage
Helpful Dataset 7000 reviews
2000 reviews
1000 reviews
BERT Helpfulness Discriminator
Fake Reviews Generator GPT-2
3000 reviews Fine tuning
1000 reviews
2000 reviews
BERT Fake reviews Discriminator
Test stage
BERT Helpfulness Discriminator
1000 reviews
BERT Fake reviews Discriminator
Source credibility
Trustworthiness
P > 0.8

Figure 1. Scheme of the methodology

2.1. Data extraction and preprocessing

The original dataset has been extracted from Ni et al. (2019), which provides many reviews from different types of products. For this work we have selected a specific dataset, *"Cell phones and accessories*" (1,128,437 reviews). The dataset has been preprocessed to filter and to keep only the useful reviews. Firstly, the reviews with no votes have been deleted. Secondly, after analysing the dataset, we found that the reviews with a date prior to 2012 generally had lower number of votes. Thus, to have a variety of values among the votes, we have filtered and kept only the reviews with a date after 2012. Finally, we have filtered again and kept only the reviews with a verified user to ensure the trustworthiness of the review. After this preprocessing, the dataset has been ordered in quartiles according to the number of votes. The reviews with two votes or below have been considered as *not helpful*. The reviews with nine votes or more have been considered as *helpful*. The final *"helpful dataset"* has been generated by extracting 3500 *not helpful* reviews and 3500 *helpful* reviews. The dataset has been used for two purposes: (1) to provide an input of *"truthful dataset"* for the fake reviews generator and (2) to provide a *"helpful dataset"* for the BERT Helpfulness Discriminator.

2.2. Fake reviews generation using GPT-2

After the preprocess of the original dataset, the fake reviews generator has been trained and used to generate a *"fake dataset"*. We have used the GPT-2 transformer model developed by

Radford et al. (2019). The GPT-2 model has been pre-trained to generate realistic texts. The model only needs to be fine-tuned to generate texts on a specific topic. We have fine-tuned the model using 1000 steps with an average training loss of 0.92 at the end of the execution. 3000 reviews from the *"helpful dataset"* that were generated in the previous step have been used as training data. Based on all of this, our customized model is able to generate fake reviews on the product category "Cell phones and accessories".

Using this model, we have generated 1000 fake reviews. Afterwards, we have created a new dataset made of the newly generated fake reviews and 1000 truthful reviews from the *"helpful dataset"*. Then, we have added a new column assigning a 0 if the review was truthful or 1 if it was fake. This dataset called *"fake-truthful dataset"* will be further used for the classifier step.

2.3. Fake and Helpfulness reviews classifiers

Once the *"helpful"* and *"fake-truthful"* datasets have been created, the next step was to train the classifiers to predict the helpful and the truthful reviews. The classifiers used in this work are based on the BERT model developed by Devlin et al. (2018). We have trained two different classifiers: one to predict helpful reviews (BERT-Helpful) and the other one to predict truthful reviews (BERT-Truthful). BERT-Helpful has been trained using 2000 reviews from the *"helpful dataset"*, which were truthful reviews combining helpful and not helpful reviews. BERT-Truthful has been trained using 2000 reviews from the *"fake-truthful"* dataset.

Once both models have been trained and validated, we have generated predictions using the same dataset for both models in the test stage. The dataset contained 1000 fake and truthful reviews, which were not used for training the classifiers. The objective was to combine both classifiers to determine the reviews that were at the same time helpful and truthful, the so-called trustworthy reviews. BERT-Helpful predicts a specific range of reviews as helpful. BERT-Truthful also predicts another range of reviews as truthful. Our goal is to merge both ranges of reviews, helpful and truthful, to find the ones contained in there.

Finally, the pool of reviews has been analyzed, characterizing the features that make a review trustworthy. This characterization has been made using *KeyBert*, developed by Grootendorst (2020). This tool allows the extraction of keywords for BERT models. It can be customised to extract keywords of different lengths starting in 1 (keywords or keyphrases) or to set the diversity of keywords we want to obtain.

3. Results

3.1 Experimental setup

The experimental evaluation has been conducted using the free version of *Google Collab*. The resources provided by the free version are a single node with one GPU Tesla K80 and 64 GB of RAM. Within this version, the availability of resources are not guaranteed and limited, and sometimes the usage limits fluctuates depending on the demand. The code has been developed using *Python3* and pandas library for the datasets management and processing. The GPT-2 transformer model has been applied with the *gpt-2-simple* library[1], which uses *TensorFlow* and wraps existing model fine-tuning and generates texts for GPT-2. It uses specifically the "small" 124M and "medium" 355M hyperparameter versions. We have used the "small" 124M version of the model for this work. Thus, the model can be handled by our system. The BERT classifier has been used with a version developed in *PyTorch*[2].

3.2 Classifiers performance

BERT-Helpful has been trained in 20 epochs with a training loss of 0.384 and a validation loss of 0.268. BERT-Truthful has been trained in 20 epochs with a training loss of 0.378 and a validation loss of 0.283. The testing for the classifiers has been made using the same dataset combining 1000 fake and truthful reviews. The accuracy is 0.9 for BERT-Helpful classifier and 0.91 for BERT-Truthful.

3.3. Trustworthy reviews extraction and analysis

Once the predictions have been generated, the next step is to merge results and extract the common reviews that both classifiers have to discover the trustworthy reviews. The default classifier uses a threshold of 0.5 to classify the data: whether it is helpful or not, or whether it is truthful or not. We wanted to ensure that our data was truly helpful and truthful. So the threshold needs to be adjusted.

Figure 2 explores different thresholds applied to the classifiers and the number of trustworthy reviews we got. It can be observed that as the threshold increases, the number of trustworthy reviews decreases proportionally. The higher the threshold, the higher the number of trustworthy reviews. We have fixed the threshold at 0.8, since at 0.9 the number of reviews decreases too much.

[1] https://github.com/minimaxir/gpt-2-simple

[2] https://github.com/prateekjoshi565/Fine-Tuning-BERT

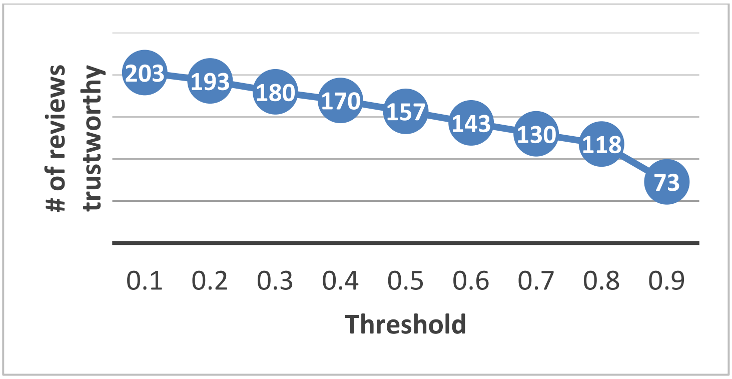

Figure 2. Exploration for different # of reviews trustworthy per threshold applied

Finally, in order to facilitate a deeper interpretability of results, we have gotten the main keywords that are most similar to a document from the 118 trustworthy reviews using the tool *KeyBert*. After doing a qualitative analysis of those extracted keywords we have identified that the content of trustworthy reviews has to do with mobile phones brands (e.g., Samsung Galaxy, iPhone), accessories brands (e.g. otterbox), the physical features of the mobile phone (e.g., long, size, sleek, weight, etc.), the physical features of the mobile phone accessories (e.g., quality of glass screen protector, adaptability of cases to the mobile phones), the technical features of the mobile phone (e.g., sound quality, mobile network coverage/signal, hardware performance) and about the UX (e.g., screen touch responsiveness).

4. Conclusions

In this work we have developed a methodology to detect and characterize trustworthy reviews within the product category “Cell phones and accessories”. We have generated fake reviews using the GPT-2 model to train a classifier to detect truthful reviews. Likewise, a second classifier has been trained to detect helpful reviews. We obtain a higher accuracy, around 0.9, for both classifiers and 118 trustworthy reviews from a dataset of 1000 reviews. Finally, the keywords allow us to characterize some of the trustworthy reviews' common patterns for this product category, thus understand the content of those reviews.

One limitation of this work was the limited resources available, thus it was necessary to reduce the number of data used to generate our results. Future work could incorporate more powerful systems able to analyse much larger datasets from different product categories. Moreover, the detection and analysis of trustworthy reviews could be improved.

References

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

Chakraborty, U. (2019). The impact of source credible online reviews on purchase intention: The mediating roles of brand equity dimensions. Journal of Research in Interactive Marketing.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Grootendorst, M. (2020). KeyBERT: Minimal keyword extraction with BERT(Version v0.3.0). Version v0.3.0. doi:10.5281/zenodo.4461265

Hsieh, J. K., & Li, Y. J. (2020). Will you ever trust the review website again? The importance of source credibility. International Journal of Electronic Commerce, 24(2), 255-275.

Joshua L. Wiener and John C. Mowen (1986) ,"Source Credibility: on the Independent Effects of Trust and Expertise", in NA - Advances in Consumer Research Volume 13, eds. Richard J. Lutz, Provo, UT : Association for Consumer Research, Pages: 306-310.

Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (2020). FNDNet–a deep convolutional neural network for fake news detection. Cognitive Systems Research, 61, 32-44.

Kelly, H. (1973), "The Process of Causal Attribution," American Psychologist, 28, 107-128.

McGinnes, E. and C. Ward (1980), "Better Liked Than Right: Trustworthiness and Expertise in Credibility,-Personality and Social Psychology Bulletin, 6, 467-472.

Metzger, M. J., Flanagin, A. J., and Medders, R. B. Social and heuristic approaches to credibility evaluation online. Journal of Communication, 60, 3, 2010, 413–439.

Ni, J., Li, J., & McAuley, J. (2019, November). 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).

Ozbay, F. A., & Alatas, B. (2020). Fake news detection within online social media using supervised artificial intelligence algorithms. Physica A: Statistical Mechanics and its Applications, 540, 123174.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Xie, H. J., Miao, L., Kuo, P. J., & Lee, B. Y. (2011). Consumers' responses to ambivalent online hotel reviews: The role of perceived source credibility and pre-decisional disposition. International Journal of Hospitality Management, 30(1), 178-183.

Shan, Y. (2016). How credible are online product reviews? The effects of self-generated and system-generated cues on source credibility evaluation. Computers in Human Behavior, 55, 633-641.

Utz, S., Kerkhof, P., & Van Den Bos, J. (2012). Consumers rule: How consumer reviews influence perceived trustworthiness of online stores. Electronic Commerce Research and Applications, 11(1), 49-58.

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management, 57(2), 102025.