Pre-Print Version of Paper Accepted for Journal of Retailing and Consumer Services

Unfolding the Characteristics of Incentivized Online Reviews

Abstract

The rapid growth of social media in the last decades led e-commerce into a new era of value co-creation between the seller and the consumer. Since there is no contact with the product, people have to rely on the description of the seller, knowing that sometimes it may be biased and not entirely true. Therefore, review systems emerged to provide more trustworthy sources of information, since customer opinions may be less biased. However, the need to control the consumers' opinion increased once sellers realized the importance of reviews and their direct impact on sales. One of the methods often used was to offer customers a specific product in exchange for an honest review. Yet, these incentivized reviews bias results and skew the overall rating of the products.

The current study uses a data mining approach to predict whether or not a new review published was incentivized based on several review features such as the overall rating, the helpfulness rate, and the review length, among others. Additionally, the model was enriched with sentiment score features of the reviews computed through the VADER algorithm. The results provide an in-depth understanding of the phenomenon by identifying the most relevant features which enable to differentiate an incentivized from a non-incentivized review, thus providing users and companies with a simple set of rules to identify reviews that are biased without any disclaimer. Such rules include the length of a review, its helpfulness rate, and the overall sentiment polarity score.

Keywords: incentivized online reviews, text mining, sentiment analysis.

Introduction

Online shopping has become a widespread form of business over the Internet. It allows consumers to buy goods or services anytime, anywhere, using devices such as computers, tablets, and smartphones. Given that there is no physical interaction with the product or seller, customers rely not only on the description of the product but also on the comments provided by other customers that also buy the same product (Mudambi and Schuff 2010; Moro et al. 2017). Product reviews are a critical factor in the decision-making process of purchasing an item online (Blazevic et al. 2013). Although at different scales, both positive and negative reviews have a direct impact on the decision of the customer to buy a product (Hu et al. 2008; Lee et al. 2008). Additionally, online reviews are also useful to sellers because they provide information on their customers' opinion regarding their products. Thus, the reliability of this information exchange format is crucial (Hajli 2018).

The current study aims to understand the biasing phenomenon previously observed and reported by Hu et al. (2011) using data from Amazon.com, one of the largest e-commerce platforms worldwide (Etzioni 2017). Companies may use "paid reviews", written by persons or companies that charges for the service. However, there are also "incentivized reviews", written by regular customers who acquired the product for free or at a discount. According to recent studies, in particular, the one conducted by ReviewMeta (2016), the overall rating of a product is influenced by incentivized reviews. Although Amazon recently started banning such reviews (Amazon.com 2016), it is naïve to assume that this practice will cease to exist. This type of reviews may continue to influence the rating system; only it is now harder to identify bias and ignore it in the process of decision making because there are no disclaimers to confirm it. Different methods for recalculating an overall rating of a product exist. However, there is a noticeable absence mechanisms

with the purpose of identifying bias in a new-coming review that has no disclaimer. Although past research has already studied cues for highlighting fake reviews using experimental studies (Munzel, 2016), the current study addresses yet another phenomenon, namely, to identify markers for predicting real but incentivized reviews. This work aims to bridge such gap by helping users of e-commerce platforms to properly assess the information on which they base their decisions on. The main contribution of this study is to create a classification model for online reviews using data and text mining techniques to find patterns and to predict the probability of a new review being biased.

The paper is organized as follows: “Literature Review” section offers a brief introduction on online review systems, incentivized reviews and their relevance, and the text mining techniques that are of interest for this paper. In the “Methodology” section, the datasets and methods used are presented. In the section “Results and Discussion” the results obtained are discussed. Finally, the “Conclusion” section presents the final comments and findings on the current paper.

Literature Review

Online reviews and e-WOM

One of the most important channels for electronic word-of-mouth (e-WOM) dissemination is online customer review systems (Dellarocas 2003). Customer reviews and comments on products or services appeal to the very human need to know "what everybody else is doing". Since high levels of trust exist in information obtained from online networks (Grabner-Kräuter 2009), reviews can move shoppers from consideration to purchase (Bulmer and DiMauro 2009). The results of a study conducted by Utz et al. (2012) aimed at examining the impact of online reviews on consumer trust in an online store showed that reviews turned out as the strongest predictor of trustworthiness

judgments. Store reputation had no significant effect compared to the reviews. Therefore, e-WOM plays an important role in consumer decision making, indicating that online consumer communities indeed empower consumers. Consumer have different motives to write reviews, from ego involvement to subjective norms and sometimes even the need for vengeance (Dixit et al., 2018). However, some consumers may also be driven to write a review in the form of paid or incentivized reviews. Paid reviews are bought as a service from companies whose job is to create positive reviews about a product to increase its sales. Incentivized reviews, on the other hand, are written by consumers who acquired the product for free or at a big discount from the seller in exchange for an "honest and unbiased review" – which is, nonetheless, still a form of paid review (Petrescu et al. 2018).

This study uses data from one of the major e-commerce companies worldwide - Amazon. Amazon's review system is an important asset of its business. However, on this platform, there are a lot of incentived reviews. Although the reviewers who write these reviews claim they express their real opinion on the product – positive or negative – these incentivized reviews tend to be overwhelmingly biased in favor of the rated product, according to the study conducted by ReviewMeta (2016). Users asked to write reviews tend to write them using explaining language, which may carry more sentiment and provide a better interpretation of the product (Moore 2012). These reviews, with strong sentiment, are more effective to consumers (Kim et al. 2016). The results from the study conducted by ReviewMeta (2016), over 7 million Amazon reviews, showed that the average rating for products with incentivized reviews was higher than non-incentivized ones (4.74 versus a 4.36 average rating, out of 5 stars). Although the difference of 0.38 is small, the impact was considerable. Based on the rating distribution also calculated in the same study, products would rise from the 54th to the 94th percentile of the rated products

if they had incentivized reviews. Indeed, incentivized reviews may create top-rated products.

Amazon recently tried to force incentivized reviews to be written only under a specific program called Vine so that such reviews could be easily classified. However, despite all efforts to have the most transparent review system, this type of reviews will always exist, particularly holding no disclaimers to prove their biased nature. There is a need to study the text contents and find patterns that may show the difference between an incentivized and a non-incentivized review without needing a badge or disclaimer to identify them. Text mining enables to do so by offering a wide variety of tools for extracting patterns of hidden knowledge from unstructured text, including sentiment analysis.

Mining online reviews

Text mining is the field of computer science research that tries to solve the crisis of unstructured information overload by combining techniques from data mining, machine learning, natural language processing, information retrieval and knowledge management (Feldman and Sanger 2007). Text mining can be traced back to Feldman and Dagan's presentation in 1995. The exponential growth of textual data as a result of the Internet and social media led to intense research on text mining. Techniques for tasks such as summarization and part-of-speech tagging have been developed on the past few decades to leverage decision making (Miner et al. 2012).

Text mining is a complex task since it involves dealing with unstructured data (Tan 1999). However, it is a worthy challenge as the majority of business-relevant information is in an unstructured format, primarily text (Linstedt 2006). Text mining techniques have already been used in several businesses, including e-commerce platforms. In a study by

Cao et al. (2011), content analysis was used to quantify the feedback in text comments. Findings suggest that rich content plays an essential role in building buyers' trust on a seller. Text mining can help an organization derive potentially valuable business insights from text-based content such as posts, comments or reviews on social media (Heng, et al. 2018; Calheiros et al. 2017; Guerreiro and Moro 2017; Guerreiro et al. 2016). Natural language processing (NLP) considers the hierarchical structure of language: several words make a phrase, several phrases make a sentence and, ultimately, sentences convey ideas. By analyzing language for its meaning, NLP systems have successfully performed tasks such as correcting grammar, converting speech to text and automatically translating between languages (Church and Rau 1995). By using NLP, it is possible to organize and structure knowledge to perform several tasks such as translation, relationship extraction, speech recognition, topic segmentation, and sentiment analysis (Humphreys and Wang, 2017; Villarroel Ordenes et al., 2017). An active research area of NLP is sentiment analysis (Cambria et al. 2015), which can be traced back to early 2000s when artificial intelligence researchers unleashed its power to understand user behavioral patterns (Nasukawa and Yi 2003). Currently, it has been applied in almost every business and social domain. Regarding social commerce, sentiment analysis has large relevance in online reviews that convey customers' opinions about a product or a service. Opinions are central to almost all human activities and represent key influencers of human behavior. Our beliefs and perceptions of reality and the choices we make are conditioned upon how others see and evaluate the world (Moro et al. 2018b). For this reason, when we need to make a decision we often seek out the opinions of others (Constantinides and Holleschovsky 2016).

Recently, a new algorithm called VADER (Hutto and Gilbert 2014) was developed to perform sentiment analysis and, according to its authors, it outperforms most of the

competitors' tools. The reason for its high performance is related to the fact that it takes into consideration several factors usually ignored, such as capitalization and an excess of punctuation, among others, that improves the accuracy of the review's sentiment score. Hutto and Gilbert (2014) began by constructing a list inspired by well-established sentiment banks like Linguistic Inquiry Word Count (LIWC), General Inquirer (GI) and Affective Norms for English Words (ANEW). Next, several lexical features common to sentiment expression in social media, such as emoticons, slang, acronyms and initialisms, were added to their list. The output of VADER is a compound score, which is a unidimensional measure of sentiment for a given sentence - it is a normalized weighted composite score. It is computed by summing the valence scores of each word in the lexicon and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive) (Hutto and Gilbert, 2014). In our study, we take advantage of such lexicon, which has proven its usefulness in other studies (e.g., Araújo et al., 2016; Ribeiro et al., 2016). The sentiment-related features extracted through VADER may enrich a data mining model in an attempt of classifying whether or not a review is incentivized.

The current paper aims to present a predictive model that may be able to accurately classify an incentivized review without any disclaimer, using the characteristics of the text. According to Moore (2012), users who are asked to write reviews tend to do it a more elaborate language, which carries more emotion and provides a better interpretation of the product for other users. This works for both positive and negative reviews. Kim et al. (2016) have confirmed such behavior through their study also based on empirical work using Amazon reviews. On the same line of thought but on the tourism context, Mayzlin et al. (2014) discovered that hotels investing on incentivized reviews have a greater share of positive reviews on TripAdvisor (anyone can write a review) relative to Expedia (users need to have booked a night). Such result suggests that the biased behavior derived from

an incentivized reviewer is context independent. Thus, if a user had a good product experience, s/he usually intends to show other users that her/his opinion is honest and not bought, and so s/he tends to be detailed about the positive aspects of the product. On the other hand, if the experience was not so good and the users still have to write their feedback, they will be more extra explanatory so they can excuse themselves in the eyes of the seller, who gave them something. Either way, the users feel like they have to justify their opinion. We expect that our model may use length of the review and sentiment of the review to classify opinions such as:

> **H1**: Incentivized reviewers tend to be more extensive in their reviews than regular reviewers.
>
> **H2**: Incentivized reviewers tend to express more positive sentiments in their reviews than regular reviewers.

Additionally, to improve model's accuracy, we included also known relevant features on a broader scope of online reviews. The overall score is a known construct, widely studied in several online reviews' contexts (e.g., Hu et al. 2008), and recently Petrescu et al. (2018) found that it is also influenced by the fact that a reviewer has received an incentive to write the review. Thus, we argue that the effect is likely the same of the sentiment score, in alignment with H2 (thus, incentivized reviewers tend to grant higher scores). Another feature included is the number of helpful votes a review has received. Some authors found a relation between the number of helpful votes with the overall granted score (e.g., Silva et al. 2018). Such finding suggests that the helpful votes can also help in classifying incentivized reviews.

Methodology

The database used was retrieved from the University of California San Diego and first analyzed by McAuley et al. (2015; 2016). It consists of more than 140 million product reviews and metadata from Amazon, spanning from May 1996 to July 2014. The dataset includes review features such as ratings, text and helpfulness votes, product metadata features (e.g., descriptions, category information, price, brand and image features), and links from other also viewed/also bought related products. The dataset is separated into 24 categories including Books, Electronics, Movies and TV, among others. The database also includes a collection of 5-core subsets that holds reviews from users and items that have at least five reviews each. To avoid sparse data, this study collected the samples from 5-core subsets.

Data preparation

For the current study, books and the electronics categories were selected. The datasets represent the two largest subsets of the main dataset: 8,898,041 reviews for books and 1,689,188 for electronics. These two subsets are unbalanced, i.e., most of the consumers opinions are non-incentivized reviews. To address such problem, a random balanced sample was generated from each one (i.e., with a similar number of both incentivized and non-incentivized reviews). The extraction task was performed by a set of rules written in Python, built specifically for this step. The samples were extracted in two steps: the first run through the data extracted all the incentivized reviews found in the 5-core subset, while a second run extracted a randomly selected set with a similar number of non-incentivized reviews. A review was classified as incentivized if it contained any disclaimer in its text. After an exploratory analysis of hundreds of reviews, a list of the observed disclaimers was created. The list also included some other variations of these terms. In the first run through the data, a case insensitive search was performed over all reviews to detect those containing at least one of the disclaimers in the above mentioned

list. After verification of the extracted reviews, we double-checked a randomly selected sample to assess the suitability of those reviews as incentivized ones. All the four authors with different expertizes (i.e., consumer behavior, digital marketing, e-commerce) assessed this sample. As a result, more than 40 terms were excluded. Table 1 shows the final list of 26 expressions used as disclaimers.

Table 1 – List of disclaimer expressions

disclaimer	discount for review
discount to review	for the purpose of a review
free for my review	free for review
free reviewer's sample	free sample
free to review	Freebie
in exchange for a review	in exchange for my honest
in exchange of a review	in return for a review
in return of a review	product for review
product for test	review for product
review sample	review unit
reviewing purposes	sample for an honest review
sample for review	sent this for review
testing and review purposes	product sent for review

After extracting the subset of incentivized reviews, the challenge was to identify a similar sized subset of reviews that were clearly non-incentivized. Thus, we first needed to find reviews in which the reviewer had clearly mentioned a purchase intention, to target also experienced reviewers as it is the case of incentivized reviewers (du Plessis and Dubois 2015). Although many expressions may indicate a purchase, given that our sets were very large, and we only needed to find a set with a similar size to the one with incentivized reviews, we considered the expression “I bought”. As for identifying the disclaimer expressions of incentivized reviews, we performed an exploratory analysis of hundreds of reviews to make sure that such manageable sample did not contain expressions that could denote an incentivized review. Thus, the set of non-incentivized reviews was collected based on the assumption that it could not have any expression from Table 1 and also that the reviewer had actually bought the product. However, a known limitation is

that this procedure excludes other reviews where reviewers wrote more subtle messages. As du Plessis and Dubois (2015) argue, a more explicit expression may be traced to a more experienced reviewer, suggesting that our sample encompasses reviews written by experienced reviewers. Yet, as incentivized reviewers are also experienced, the abovementioned limitation turns out to target our analysis to the segment of experienced reviewers. Thereafter, the program ran a second time, stopping once it extracted the same number of cases as for the incentivized reviews subset. Therefore, the samples generated were balanced by having 50% of incentivized reviews and 50% of non-incentivized reviews. An imbalanced dataset is known to affect modeling performance, resulting in worse classifications (Chawla et al. 2004). Thus, balancing data constitutes a valid approach to address this issue (He and Garcia 2008). A random model hypothetically correctly classifies 50% of the reviews. Depending on the problems, a good model can accurately classify more than 70% (Moro et al. 2014).

Reviews that had at least one vote from other consumers were selected and the final books dataset consisted in a total of 105,202 reviews, while the electronics dataset had 5,594 reviews in total (half incentivized and half non-incentivized for both datasets).

The sentiment analysis was performed using the VADER algorithm (Hutto and Gilbert 2014) on two different levels for each review in the datasets. First, the sentiment scores of the review are computed. These scores describe the sentiment intensity of the entire review and present how positive or negative a given review is. Secondly, the sentiment analysis was carried out on a sentence level. Besides the sentiment scores, calculated through VADER three length-related variables were also included: number of characters, number of words and number of sentences in a review and also a flag to indicate whether a review was incentivized or not.

Table 2 shows a more detailed explanation of the variables that constitute the final dataset.

Table 2 – Variables Description

Variable	Mean-SD Books	Mean-SD Electronics	Description
Reviewed	-	-	ID of a review. Assigned number during the extraction process to easily access the text of a review.
reviewerID	-	-	Amazon reviewer ID, provided in the original dataset.
Asin	-	-	Amazon product ID, provided in the original dataset.
unixReviewTime	-	-	Time when the review was written in unix format, provided in the original dataset. It places a review in a specific moment in time and for that reason it will be excluded from the model, since it cannot be used as a predictive variable to classify new reviews.
reviewTime	-	-	Date when the review was written, in raw format, provided in the original dataset.
number_chars_ review	1319.7(1298.5)	1585.1(1844.3)	Number of characters in the review text. Calculated in the extraction process.
number_words_ review	234.6(224.1)	285.5(325.5)	Number of words in the review text. Calculated in the extraction process.
number_ sentences_review	11.1(10.1)	13.0(14.3)	Number of sentences in the review text. Calculated in the extraction process.
Overall	4.1(1.2)	4.2(1.1)	Star rating given by the reviewer to the product, scored from 1 to 5, provided in the original dataset.
Helpful	0.74(0.33)	0.81(0.29)	Helpfulness rate of a review measures the proportion of helpful votes on the total votes. The mean calculation shows the mean of the ratio between helpful_votes and total_votes when there is at least one total vote. In the original dataset, 2 values are provided: number of helpful votes and number of total votes (e.g. 2/3, which means that 3 people voted that review: 2 found it helpful and the other one did not). In order to have a normalized score, this parameter was converted in a scale from 0 to 1.
helpful_votes	4.11(45.66)	10.03(52.20)	Number of helpful votes of a review, provided in the original dataset.
total_ votes	5.40(49.98)	11.19(55.24)	Number of total votes of a review (helpful and not helpful), provided in the original dataset.
overall_positive_ value	0.17(0.77)	0,13(0.64)	Positive score of the review as a whole, calculated through the VADER algorithm.
overall_negative_ value	0.05(0.46)	0.04(0.37)	Negative score of the review as a whole, calculated through the VADER algorithm.
overall_neutral_ value	0.78(0.76)	0.83(0.66)	Neutral score of the review as a whole, calculated through the VADER algorithm.

overall_ compound_value	0.72(0.51)	0.71(0.49)	Compound score of the review as a whole, calculated through the VADER algorithm. This is a single measure of polarity of the review, ranging in a scale from -1 (extremely negative) to 1 (extremely positive).
number_positive_ sentences	0.43(0.81)	0.34(0.73)	Number of positive sentences in the review. This parameter is calculated running the algorithm one sentence at a time. If the positive score is the highest out of the three, the sentence is classified as positive.
number_negative _sentences	0.11(0.40)	0.08(0.33)	Number of negative sentences in the review. This parameter is calculated running the algorithm one sentence at a time. If the negative score is the highest out of the three, the sentence is classified as negative.
number_neutral_ sentences	10.52(9.66)	12.57(13.86)	Number of neutral sentences in the review. This parameter is calculated running the algorithm one sentence at a time. If the neutral score is the highest out of the three, the sentence is classified as neutral.
avg_positive_ sentences	0.19(0.30)	0.16(0.29)	Average of the positive scores of the positive sentences, calculated through VADER algorithm. If there are none in the review, this parameter is null.
avg_negative_ sentences	0.05(0.18)	0.05(0.18)	Average of the negative scores of the negative sentences, calculated through VADER algorithm. If there are none in the review, this parameter is null.
avg_neutral_ sentences	0.81(0.71)	0.85(0.61)	Average of the neutral scores of the neutral sentences, calculated through VADER algorithm. If there are none in the review, this parameter is null.
compound_ sentences_ average	0.26(0.22)	0.22(0.19)	Average of the compound scores of all the sentences on a review, calculated through VADER algorithm.
summary_ positive	0.34(0.33)	0.34(0.30)	Positive score of the review summary, calculated through the VADER algorithm.
summary_ negative	0.07(0.18)	0.05(0.15)	Negative score of the review summary, calculated through the VADER algorithm
summary_ neutral	0.60(0.33)	0.61(0.30)	Neutral score of the review summary, calculated through the VADER algorithm.
Incentivized	0=52601 1=52601	0=2797 1=2797	Categorical variable that labels the review as being incentivized (1) or non-incentivized (0).

Modeling

The purpose of the current study is to predict whether a non-disclosed review was incentivized or not; therefore a classification model was used. For this type of problems, the most common models used are decision trees (Mitchell, 1997), random forests (Liu

et al., 2013), neural networks (Bishop, 1995), support vector machines (Burges, 1998) and Bayesian networks (Darwiche, 2010). Despite the accuracy of support vector machines (SVM) and neural networks (NN) to model complex non-linear relationships (e.g., Moro et al. 2014), they usually lack the advantages of decision trees (DT) regarding interpretability: users can easily understand the rules behind the decisions based on such models. The divide-and-conquer strategy of decision trees provides easy-to-follow decisions and the relevance of easily understanding each feature's influence (Moro et al. 2018a). Therefore, DT are used in the current study. Random forests (RF) consist in ensembles of decision trees, where each tree contributes individually to the model's overall performance (Gashler et al. 2008). Nevertheless, RF have the same limitation of NN and SVM since they cannot be directly interpreted, thus favoring accuracy over interpretability. Considering that recent studies in different contexts have shown RF outperform techniques such as neural networks and support vector machines (e.g., Liu et al. 2013; Naghibi et al. 2017), we have also adopted RF for comparison purposes.

Three DT models were generated for each dataset to compare their performance based on two different decision tree algorithms - C5.0 (Quinlan 1993) and C&RT (Brieman et al. 1984), and an additional RF model. The main differences of DT techniques rely on the algorithms used. C5.0 uses information gain maximization, and entropy reduction and C&RT uses the GINI index for splitting the tree. However, they are both usually used successfully for classification purposes (Wu et al. 2008). The DTs adopted for our experiments use split selection methods that perform feature selection by using a top-down recursive divide-and-conquer technique (Ratanamahatana and Gunopulos, 2003). Such a selection chooses the attributes that maximize information (either by using entropy-based or Gini-based measures) for classification purposes, thus allowing for predictors to be hierarchically organized until a stopping criterion is reached. DTs also

use pruning techniques to balance classification accuracy while reducing overfitting. Therefore, all the variables were used as inputs for DTs as the final model only presents those that better represent the problem at hand.

The dataset was split into a train (70%) and a test (30%) sample due to its effectiveness in building the best classification models (Sarkar, 2016). Performance measures were calculated for the three models in each dataset for comparison purposes. Table 3 shows the performance metrics of accuracy, precision sensitivity, F-measure, and specificity.

Table 3 – Performance Metrics

		Accuracy		Precision		Recall/ sensitivity		F-measure		Specificity	
		Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Electronics	C5.0	77.9%	76.6%	76.6%	78.3%	76.8%	76.7%	76.7%	77.5%	77.6%	76.5%
	C&RT	75.8%	75.8%	73.9%	76.8%	78.2%	77.3%	76.0%	77.0%	73.5%	74.1%
	RF	83.2%	76.0%	83.7%	76.2%	82.5%	75.8%	83.1%	76.0%	84.0%	76.3%
Books	C5.0	77.3%	75.0%	76.8%	74.8%	78.2%	76.0%	77.5%	75.4%	76.5%	74.0%
	C&RT	74.4%	74.2%	74.2%	74.2%	74.5%	74.6%	74.4%	74.4%	74.2%	73.8%
	RF	75.9%	75.1%	76.5%	75.6%	74.7%	74.0%	75.6%	74.8%	77.1%	76.2%

It is possible to see that the three techniques achieved very close results regarding performance. Apart from C&RT for Books dataset, they all correctly classified over 75% of the data. F-measure is also a good indicator of model fit and shows good results for the models presented. For every DT model generated, the training and testing metrics were very similar, which means there was no overfitting. However, RF shows a clear overfitting in the smaller electronics dataset (Fox et al. 2017). Although RF is a more complex technique than DT, the results for our models show only a slight improvement for the larger books dataset in a few metrics such as specificity and precision. As such, and given the interpretability advantage of DT, the RF was not adopted for knowledge extraction.

For the electronics dataset, C5.0 was better than C&RT and adequately categorized 77.9% of the training data and 76.6% of the testing data. Likewise, the most accurate model was also the C5.0 model for books (77.3% accuracy on the training set and 75.0% accuracy on the test set). Thus, we now discuss on how the input features contribute to each of the best achieved models (C5.0 for both cases).

Next, we highlight each individual feature's relevance to classifying an incentivized review in two horizontal bar plots (one for electronics and another for books), in a scale from 0 to 1 of relative relevance, thus, where 1 is the sum of all variables' importance. Each feature's relevance is drawn based on its contribution to the C5.0 DT (Silva et al. 2018). The most relevant variables for electronics were related to the length of the review, followed by the number of helpful votes and the average compound score of the sentences. The two length related variables were more important to determine the classification than all the others, summing up to 0.70. Figure 1 shows all the relevant variables used in the C5.0 model for the electronics dataset.

Figure 1 - Predictor variables - C5.0 model - Electronics

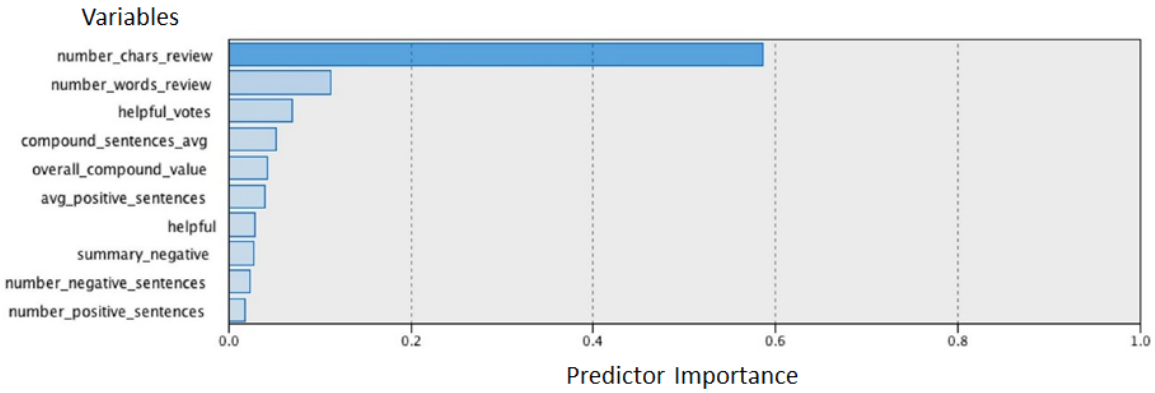
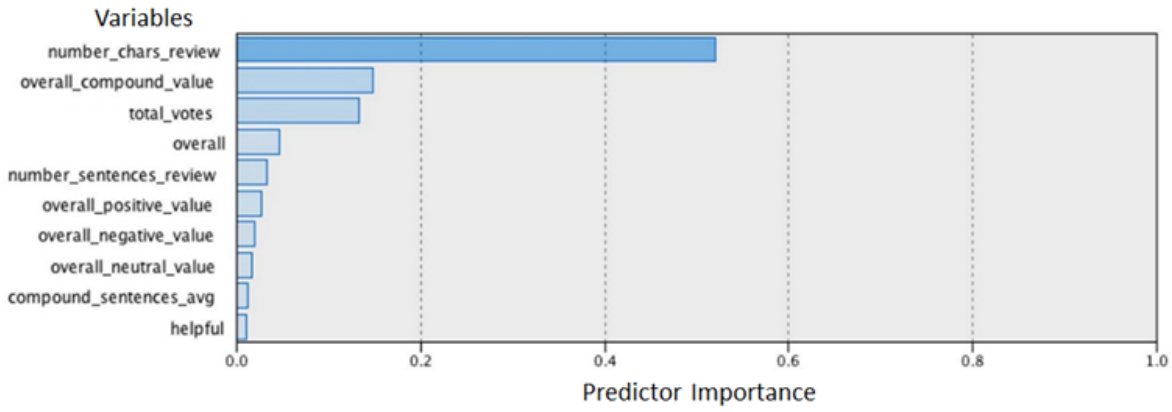


Figure 2 shows that the three most important predictor variables in this model for the books dataset are the number of characters, the overall compound value and the total number of votes, which summed up to 0.80 (i.e., the three variables combined contribute

to around 80% of the model). The relatively lower relevance of sentiment score features when compared to review length can be justified by the classification procedure described on the data preparation section. Since reviewers for both incentivized and non-incentivized are experienced ones, their opinions are likely more objective, thus with fewer emotional charge, limiting the classifier capabilities of using such information. This is especially true for the smaller set of electronics (Figure 1).

Figure 2 - Predictor variables importance - Books



Another illustrative measurement to assess the performance of a model is the ROC (Receiver Operating Characteristic) curve. The area under the curve is a known important metric for classifiers (Moro et al. 2014). An AUC value of 0.5 indicates no discriminative value (i.e., 50% sensitive and 50% specific) whereas an AUC value of 1 indicates a perfect fit model. The C5.0 model for the Electronics dataset achieved an AUC value of 0.82 for both training and testing data, which represents a fair model (e.g., Moro et al. 2014). As for the books dataset, the AUC was of 0.82 for the training set and 0.81 for the testing data.

The model was also evaluated on the unbalanced datasets to assess the robustness of each model for the real-world proportion of incentivized versus non-incentivized reviews. Table 4 shows the results for both books and electronics datasets. By comparing the results with the ones shown on Table 3, it is possible to observe a decrease in performance

for some metrics (i.e., precision and sensitivity), although accuracy remains similar to the balanced dataset model. This decrease in performance is higher for the incentivized electronics reviews than for the incentivized books reviews. However, the electronics dataset is composed by only 1.57% of incentivized reviews from the total electronics reviews, while the books dataset has 26.24% of incentivized reviews in proportion to the whole book reviews. Classification problems with rare classes (such as the case of electronics data) may suffer from such decrease in performance (Guermazi *et al*, 2018). Yet, both models are still better than a random classifier, and thus the results are useful in understanding the incentivized review phenomenon, albeit its limitations in generalization.

Table 4 – Models' evaluation on the unbalanced dataset.

Metrics	**Books**	**Electronics**
Accuracy	72.1%	78.5%
Precision	68.1%	62.7%
Recall/sensitivity	78.7%	64.9%
F-measure	73.0%	63.8%
Specificity	69.7%	78.7%

Results and Discussion

Results show that on average, a non-incentivized review has 865 characters while an incentivized has 2306, thus it is three times lengthier. Regarding word count, results are similar: the average for a non-incentivized review is 159 words, whereas an incentivized has an average of 412 words. Finally, as for the number of sentences in a review, the average in an incentivized review is 18, which is more than the double of the non-incentivized reviews – 8 sentences. A Kruskal-Wallis test showed that there is a statistically significant difference in the number of chars in a review between incentivized and non-incentivized reviews for the electronic dataset, $\chi 2(1) = 1587.737$, $p < 0.01$, with

a mean rank number of chars of 3658 for incentivized reviews and 1937 for non-incentivized. The Kruskal-Wallis test also showed that there was a statistically significant difference for Books dataset, $\chi 2(1) = 20.287{,}559$, $p < 0.01$, with a mean rank number of chars of 65938 for incentivized reviews and 39265 for non-incentivized.

In light of this analysis, H1 is supported.

To verify H2, concerning the emotional strength in a review, the overall compound sentiment score was used. To have a better understanding of the differences in sentiment scores between incentivized and non-incentivized reviews, a box plot graph was created for the overall compound scores for each dataset (books and electronics). By analyzing the graphs in Figures 5 and 6, it is possible to see that almost the whole incentivized subset, both in Books and Electronics, scored just as high as only the top half of the non-incentivized reviews.

Figure 5 - Box plot overall compound score - Electronics

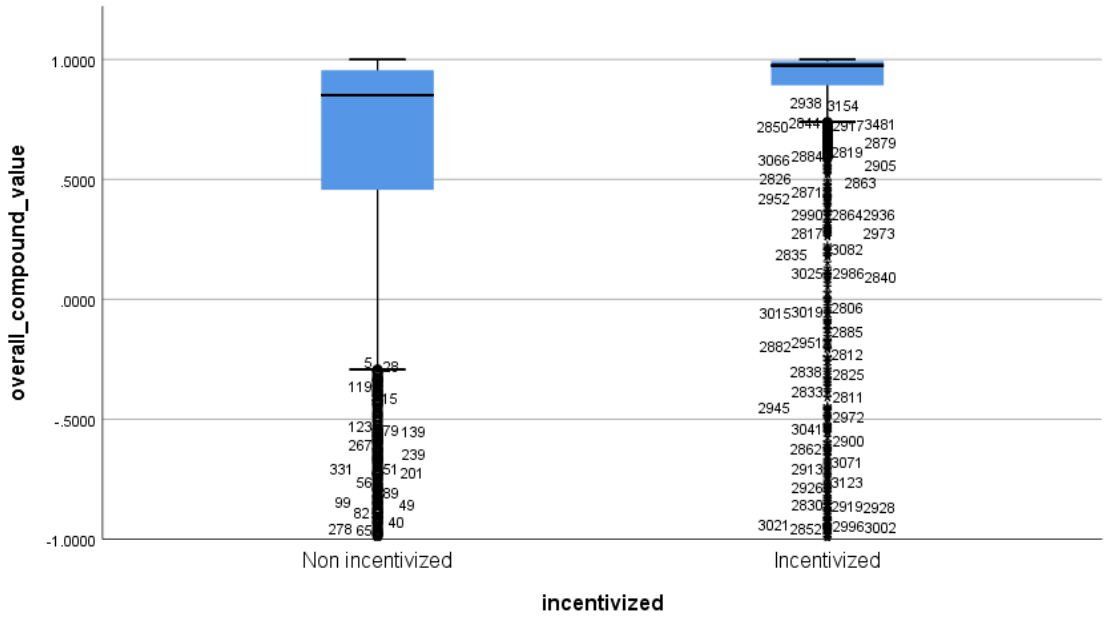
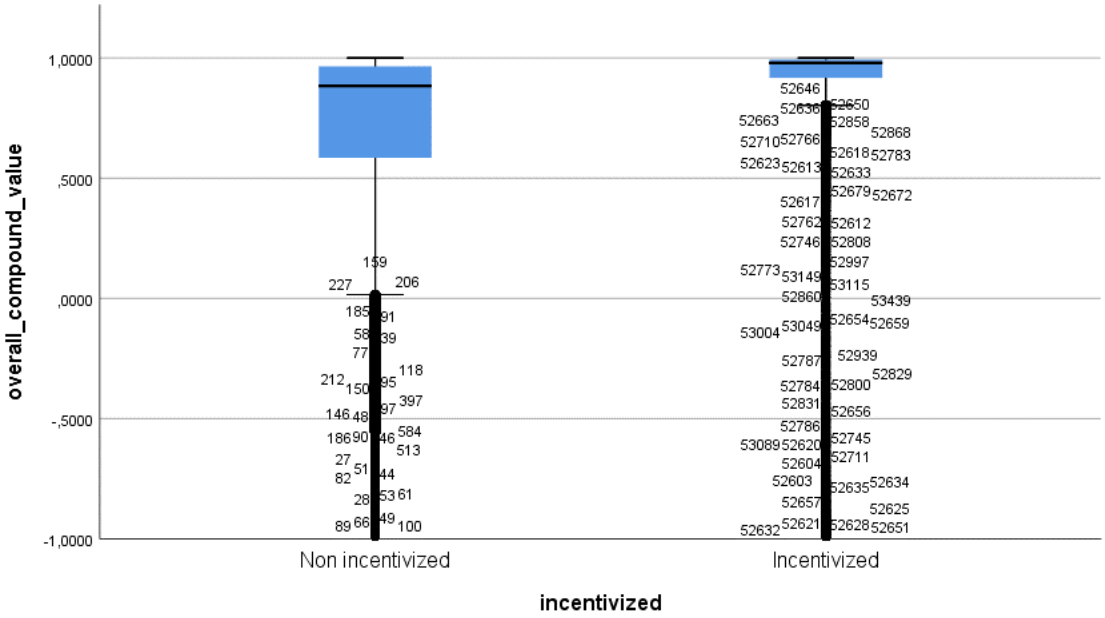


Figure 6 - Box plot overall compound score - Books

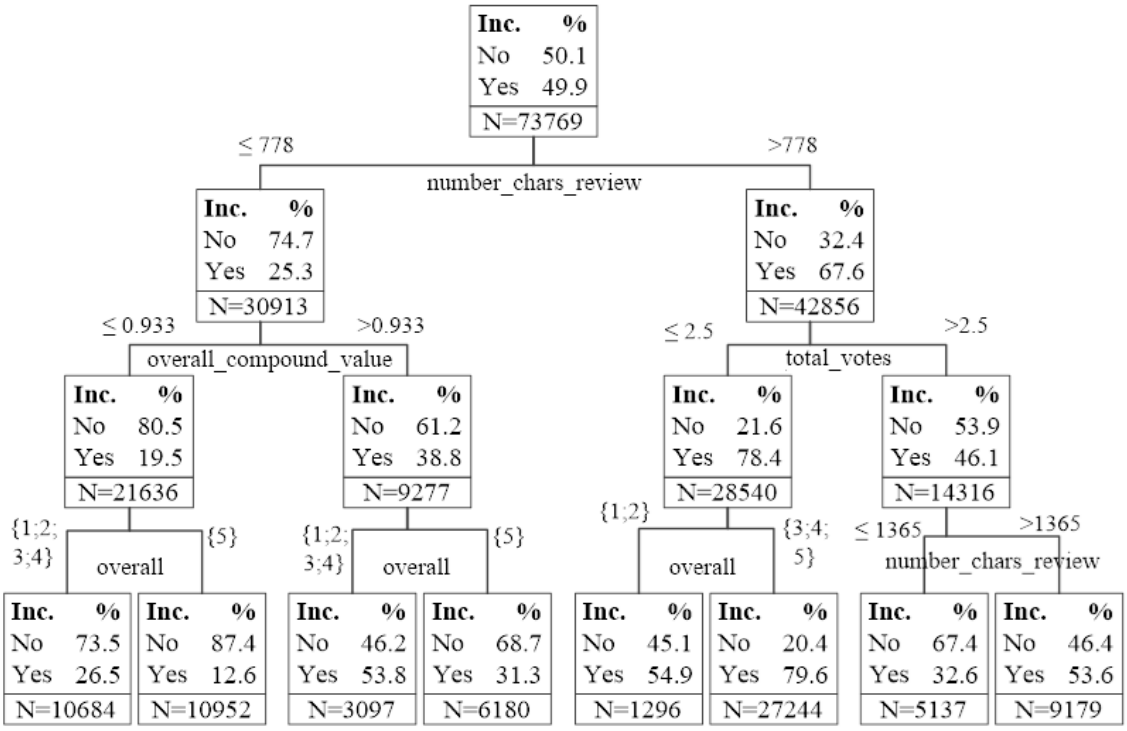


A Kruskal-Wallis test showed that there is a statistically significant difference between incentivized and non-incentivized reviews, $\chi 2(1) = 879.879$, $p < 0.01$, with a mean rank overall compound value of 3438 for incentivized reviews and 2157 for non-incentivized. The same test for the books dataset showed similar results: $\chi 2(1) = 15037.449$, $p < 0.01$, with a mean rank overall compound value of 64083.31 for incentivized reviews and 41119.69 for non-incentivized. Results support H2.

Results from the decision tree show the main criteria used to discriminate between incentivized and non-incentivized reviews is the length of the review. After that, based on the number of characters, the model either checks the number of helpful votes and the helpfulness rate or the overall compound value and the overall score of the review. An example of a rule to classify a review as incentivized is if the review has more than 778 characters, two or less votes, and an overall score of 3.0 or higher, then 79% of such reviews are labeled as being incentivized. On the other hand, if a review has 778 characters or less and the overall compound sentiment score, is lower than 0.93 it is very likely that the review is not incentivized, regardless of the overall rating. However, if the

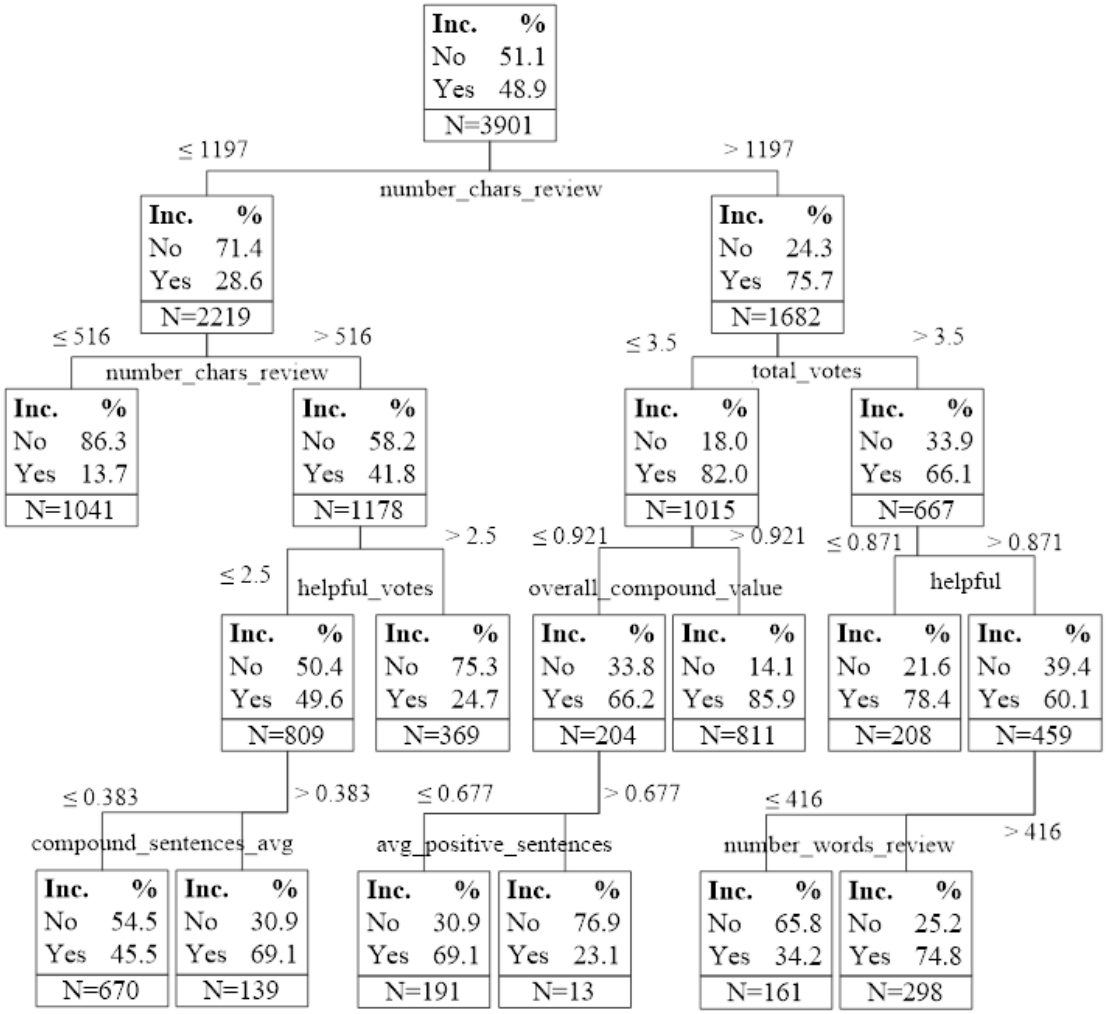
overall compound sentiment score is higher than 0.93, then the model will check the overall rating of the review, to determine whether a review should be classified as incentivized or not. Figure 7 shows the decision tree for the books dataset.

Figure 7 - C5.0 decision tree – Books (Inc.= incentivized)



According to the electronics decision tree, the first criterion used to split the tree was also the length of the review. Then, depending on the number of characters, the model checks attributes like the number of total votes, the helpfulness rate or the overall compound value. An example of a rule to classify an Electronics review as incentivized is: if the review has more than 1197 characters and less than four total votes, but the overall compound value is higher than 0.921, then 86% of such reviews are incentivized. In this dataset, there is also a very determinant rule for non-incentivized reviews: if the review has less than 517 characters, then 86% of such reviews are categorized as non-incentivized regardless of any other variable. The C5.0 decision tree for this dataset is presented in Figure 8.

Figure 8 - C5.0 decision tree – Electronics (Inc.= incentivized)



The decision tree models presented in the current paper also support H1 that states that incentivized reviews are lengthier. This is evident just by looking at the decision trees, considering the number of characters is the most important variable in both. Although the boundary numbers are different for the two datasets, it is clear to see that the higher the number, the higher the probability of the review being categorized as incentivized. For books' reviews, this number is a little lower than for electronics', but still complies with what Moore (2012) stated about incentivized reviewers tending to be more explanatory and therefore more extensive in their reviews.

For electronics reviews with a number of characters in a range from 516 to 1,197, a higher number of helpful votes is a good predictor of non-incentivized reviews. However, if there are less than three helpful votes, then the variable that determines whether or not the review is incentivized is the compound average score, where a value higher than 3.8 means that the review is most likely incentivized. In the books dataset decision tree, there is a rule showing that a review with less than 778 characters, but with a high compound score (above 0.93) and a rating of 5.0 stars is probably non-incentivized.

Considering these results, it is possible to conclude that the most important variables in predicting bias in a review are mainly structured ones, like the length of the review or the rating. One interesting fact that stands out on both datasets is the strong relationship between helpfulness of the review and incentivized reviews, which suggest that incentivized reviewers may provide more useful information because they have the motivation to write a lengthier and more detailed review. Although incentivized reviews may bias consumer decision if not properly disclaimed, such finding shows there may be a positive side to incentivized reviews in terms of their usefulness.

Conclusions

When shopping online, product reviews are known to be an important deciding factor for any potential customer. Since there is no interaction with the product itself, these reviews are strong influencers because they reflect real experiences reported by customers who acquired the item in question. Therefore, it is one of the key factors that sellers need to take into consideration. By processing this information on a regular basis, companies will be able to start acting more efficiently, which not only makes the customers happier, but it also prevents waste of money in unfocused campaigns or improvements. As previously

mentioned, some sellers on Amazon took advantage of this review tool to favor their low rated, or not rated at all, products with the purpose of increasing its sales, by actually controlling the reviews.

The current paper contributes to the literature by classifying reviews and finding patterns in the behavior of incentivized reviewers including not only sentiment score, but also other predictors related to review length and helpfulness, with the purpose of predicting bias in new-coming reviews, even if there is no disclaimer. According to the analysis performed, it was possible to conclude that incentivized reviewers do write lengthier and more sentiment charged reviews. The decision tree models generated were able to correctly predict bias in a review over 75% of times, based on some characteristics like the length of a review, the helpfulness rate and the sentiment polarity scores calculated through VADER. The most important variable, in both cases, was the number of characters of the review, which was related to one of the hypotheses tested. The most important sentiment-related variable was the overall compound score, which was higher on the incentivized reviews. Looking at the decision trees rules, it is possible to infer several decision rules to classify new incoming reviews. Finally, the contribution of this study is highlighted by the fact that incentivized reviews were recently banned by Amazon. As this type of reviews will most likely be active in forthcoming years, the current study sets the roots for identifying possible incentivized reviews not holding a disclaimer.

One of the limitations of this study is the assumption that the disclaimers used to identify incentivized and non-incentivized reviews is enough to perform a reliable extraction. It is possible that some of the observations on the datasets were misclassified on the extraction process, but considering the size of the dataset and the consequent impossibility of manually checking every review, this was considered an acceptable limitation. Another

limitation is related to the fact that only reviews from two categories of products were used. However, the categories with a higher number of reviews were used to increase the relevance of the results for the most popular items sold. There are several techniques under the text mining umbrella that can offer insightful features such as the review's underlying latent topic. Such acknowledgement justifies suggesting enriching the dataset through other text mining-based features as a future avenue of research. Additionally, since our review selection approach has arguably been focused on a set of more experienced reviewers, the presented results are only valid under such context. To address such limitation, we recommend future research based on primary data to confirm the current findings.

References

Aghaei, S., Nematbakhsh, M. A., and Farsani, H. K. (2012). Evolution of the World Wide Web: From WEB 1.0 TO WEB 4.0. International Journal of Web & Semantic Technology, 3(1), 1-10.

Amazon.com (2016), Update on Customer Reviews. Retrieved October, 2017, from https://www.amazon.com/p/ feature/abpto3jt7fhb5oc.

Araujo, M., Reis, J., Pereira, A., and Benevenuto, F. (2016). An evaluation of machine translation for multilingual sentence-level sentiment analysis. In Proceedings of the 31st Annual ACM Symposium on Applied Computing (pp. 1140-1145). ACM.

Bishop, C. M. (1995). Neural networks for pattern recognition. Oxford university press.

Blazevic, V., Hammedi, W., Garnefeld, I., Rust, R. T., Keiningham, T., Andreassen, T., Carl, W. (2013). Beyond traditional word-of-mouth: an expanded model of customer-driven influence. Journal of Service Management, 24(3), 294-313.

Brieman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. Belmont (CA): Wadsworth. Google Scholar.

Bulmer, D., and DiMauro, V. (2009). The new symbiosis of professional networks: Social media's impact on business and decision-making. Society for New Communications Research. SNCR Press.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2), 121-167.

Cao, Q., Duan, W., and Gan, Q. (2011). Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. Decision Support Systems, 50(2), 511-521.

Calheiros, A. C., Moro, S., and Rita, P. (2017). Sentiment Classification of Consumer-Generated Online Reviews Using Topic Modeling. Journal of Hospitality Marketing & Management, 26(7), 675-693.

Cambria, E., Fu, J., Bisio, F., and Poria, S. (2015, January). AffectiveSpace 2: Enabling Affective Intuition for Concept-Level Sentiment Analysis. In AAAI (pp. 508-514).

Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004). Special issue on learning from imbalanced data sets. ACM Sigkdd Explorations Newsletter, 6(1), 1-6.

Church, K. W., and Rau, L. F. (1995). Commercial applications of natural language processing. Communications of the ACM, 38(11), 71-79.

Constantinides, E., and Fountain, S. J. (2008). Web 2.0: Conceptual foundations and marketing issues. Journal of Direct, Data and Digital Marketing Practice, 9(3), 231-244.

Constantinides, E., and Holleschovsky, N. I. (2016). Impact of online product reviews on purchasing decisions. In T. A. Majchrzak, P. Traverso, V. Monfort, & K-H. Krempels (Eds.), Proceedings of the 12th International Conference on Web Information Systems and Technologies (pp. 271-278). Rome: SCITEPRESS.

Darwiche, A. (2010). Bayesian networks. *Communications of the ACM*, 53(12), 80-90.

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. Management Science, 49(10), 1407-1424.

Dixit, S., Badgaiyan, A. J., & Khare, A. (2018). An integrated model for predicting consumer's intention to write online reviews. Journal of Retailing and Consumer Services (in press). https://doi.org/10.1016/j.jretconser.2017.10.001

du Plessis, C., & Dubois, D. (2015). When and why paid reviews are bad investments: The impact of monetary incentives on reviewer certainty. ACR North American Advances.

Etzioni, A. (2017). Cyber Trust. Journal of Business Ethics, DOI: 10.1007/s10551-017-3627-y.

Feldman, R., and Dagan, I. (1995). Knowledge Discovery in Textual Databases (KDT). In KDD (Vol. 95, pp. 112-117).

Feldman, R., and Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press.

Fournier, S., and Avery, J. (2011). The uninvited brand. Business Horizons, 54(3), 193-207.

Fox, E. W., Hill, R. A., Leibowitz, S. G., Olsen, A. R., Thornbrugh, D. J., and Weber, M. H. (2017). Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. Environmental monitoring and assessment, 189(7), 316.

Grabner-Kräuter, S. (2009). Web 2.0 social networks: the role of trust. Journal of Business Ethics, 90(4), 505-522.

Gashler, M., Giraud-Carrier, C., and Martinez, T. (2008). Decision tree ensemble: Small heterogeneous is better than large homogeneous. In Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on (pp. 900-905). IEEE.

Guerreiro, J., and Moro, S. (2017). Are Yelp's tips helpful in building influential consumers?. Tourism Management Perspectives, 24, 151-154.

Guerreiro, J., Rita, P., and Trigueiros, D. (2016). A text mining-based review of cause-related marketing literature. Journal of Business Ethics, 139(1), 111-128.

Guermazi, R., Chaabane, I., & Hammami, M. (2018). AECID: Asymmetric entropy for classifying imbalanced data. Information Sciences, 467, 373-397.

Hajli, N. (2018). Ethical environment in the online communities by information credibility: a social media perspective. Journal of Business Ethics, 149(4), 799-810.

He, H., and Garcia, E. A. (2008). Learning from imbalanced data. IEEE Transactions on Knowledge & Data Engineering, (9), 1263-1284.

Heng, Y., Gao, Z., Jiang, Y., & Chen, X. (2018). Exploring hidden factors behind online food shopping from Amazon reviews: A topic mining approach. Journal of Retailing and Consumer Services, 42, 161-168.

Hu, N., Liu, L., and Sambamurthy, V. (2011). Fraud detection in online consumer reviews. Decision Support Systems, 50(3), 614-626.

Hu, N., Liu, L., and Zhang, J. J. (2008). Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. Information Technology and Management, 9(3), 201-214

Humphreys, A., & Jen-Hui Wang, R. (2017). Automated Text Analysis for Consumer Research. Journal of Consumer Research, 44(6), 1274-1306.

Hutto, C. J. and Gilbert, E. E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Kim, J., Naylor, G., Sivadas, E., and Sugumaran, V. (2016). The unrealized value of incentivized eWOM recommendations. Marketing Letters, 27(3), 411-421.

King, R. A., Racherla, P., and Bush, V. D. (2014). What we know and don't know about online word-of-mouth: A review and synthesis of the literature. Journal of Interactive Marketing, 28(3), 167-183.

Lallana, E., Quimbo, R., and Andam, Z. R. (2000). An Introduction to eCommerce: definition adapted and expanded. Philippines: DAI-AGILE, 17.

Lee, J., Park, D. H., and Han, I. (2008). The effect of negative online consumer reviews on product attitude: An information processing view. Electronic Commerce Research and Applications, 7(3), 341-352.

Liu, M., Wang, M., Wang, J., and Li, D. (2013). Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar. Sensors and Actuators B: Chemical, 177, 970-980.

Mayzlin, D., Dover, Y., and Chevalier, J. (2014). Promotional reviews: An empirical investigation of online review manipulation. American Economic Review, 104(8), 2421-55.

McAuley, J., Pandey, R., and Leskovec, J. (2015). Inferring networks of substitutable and complementary products. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM.

McAuley, J., and Yang, A. (2016). Addressing complex and subjective product-related queries with customer reviews. In Proceedings of the 25th International Conference on World Wide Web (pp. 625-635). International World Wide Web Conferences Steering Committee.

Miner, G., Elder IV, J., and Hill, T. (2012). Practical text mining and statistical analysis for non-structured text data applications. Academic Press.

Mitchell, T. (1997). Decision tree learning. *Machine learning*, 414, 52-78.

Moore, S. G. (2012). Some things are better left unsaid: how word of mouth influences the storyteller. Journal of Consumer Research, 38(6), 1140-1154.

Morente-Molinera, J. A., Pérez, I. J., Chiclana, F., and Herrera-Viedma, E. (2015). A novel group decision making method to overcome the Web 2.0 challenges. In Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on (pp. 2233-2238). IEEE.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31.

Moro, S., Rita, P., and Coelho, J. (2017). Stripping customers' feedback on hotels through data mining: the case of Las Vegas Strip. Tourism Management Perspectives, 23, 41-52.

Moro, S., Cortez, P., & Rita, P. (2018a). A divide-and-conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing. Expert Systems, 35(3), e12253.

Moro, S., Rita, P., and Oliveira, C. (2018b). Factors influencing hotels' online prices. Journal of Hospitality Marketing & Management, 27(4), 443-464.

Mudambi, S. M., and Schuff, D. (2010). What makes a helpful review? A study of customer reviews on Amazon.com. MIS Quarterly, 34(1), 185-200.

Munzel, A. (2016). Assisting consumers in detecting fake reviews: The role of identity information disclosure and consensus. Journal of Retailing and Consumer Services, 32, 96-108.

Naghibi, S. A., Ahmadi, K., and Daneshi, A. (2017). Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. Water Resources Management, 31(9), 2761-2775.

Nasukawa, T., and Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the 2nd international conference on Knowledge capture (pp. 70-77). ACM.

Petrescu, M., O'Leary, K., Goldring, D., and Mrad, S. B. (2018). Incentivized reviews: Promising the moon for a few stars. Journal of Retailing and Consumer Services, 41, 288-295.

Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.

Ratanamahatana, C. A., and Gunopulos, D. (2003). Feature selection for the naive bayesian classifier using decision trees. Applied artificial intelligence, 17(5-6), 475-487.

ReviewMeta (2016). Retrieved October, 2017, from https://reviewmeta.com/blog/analysis-of-7-million-amazon-reviews-customers-who-receive-free-or-discounted-item-much-more-likely-to-write-positive-review/

Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., and Benevenuto, F. (2016). Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Science, 5(1), 1-29.

Sarkar, B. K. (2016). A case study on partitioning data for classification. International Journal of Information and Decision Sciences, 8(1), 73-91.

Silva, A. T., Moro, S., Rita, P., and Cortez, P. (2018). Unveiling the features of successful eBay smartphone sellers. Journal of Retailing and Consumer Services, 43, 311-324.

Tan, A. H. (1999). Text mining: The state of the art and the challenges. In Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases (Vol. 8, pp. 65-70).

Utz, S., Kerkhof, P., and 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.

Villarroel Ordenes, F., Ludwig, S., De Ruyter, K., Grewal, D., & Wetzels, M. (2017). Unveiling what is written in the stars: Analyzing explicit, implicit, and discourse patterns of sentiment in social media. Journal of Consumer Research, 43(6), 875-894.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan G., Ng, A., Liu, B., Yu, P., Zhou, Z-H., Steinbach, M., Hand, D. & Steinberg, D. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.