Finding deceptive book reviews in Brazilian Portuguese

Emerson Yoshiaki Okano[0000−0001−8684−9301], Verônica Letícia Oyan de Moraes, Mateus Tarcinalli Machado[0000−0003−1848−3082], and Evandro Eduardo Seron Ruiz[0000−0002−7434−897X]

Departamento de Computação e Matemática – FFCLRP,
Universidade de São Paulo,
Avenida dos Bandeirantes, 3900 - Monte Alegre, CEP 14040-901, Ribeirão Preto–SP , Brasil.
{okano700, veronica.leticia.moraes, mateusmachado, evandro}@usp.br

Abstract. User-generated reviews of products and services play a major role in customer's decisions. However, such reviews are not always reliable. The widespread use of deceptive reviews has increased interest in datasets and methods to recognize these malicious opinions automatically. In this study, we first assembled and analyzed deceptive book reviews in Brazilian Portuguese. We collected 68 truthful and 68 deceptive reviews, which amounted to 136 reviews. We experimented with the categorization of deceptive reviews by means of Support Vector Machines (SVM) and Random forest with features such as bag-of-words (BOW), the Linguistic Inquiry and Word Count dictionary, and Latent Dirichlet Allocation. Our strategy based on SVM with BOW provided an accuracy of 66.7%. We believe that an increment in the number of reviews may lead to even higher accuracy values.

Keywords: Deceptive reviews · Opinion spam · Review spam.

1 Introduction

The web constitutes a vast source of textual information, but not every piece of information presented therein is reliable. Some users might use the web to spread false information with the purpose of manipulating and deceiving other users [10]. Most online shoppers rely on product reviews before buying a product. Unfortunately, there are people who take advantage of this behavior and create false product reviews, which are usually designated "deceptive review". These kinds of reviews are known to be an effective component of the buyer's final decision. In this paper, we consider "deceptive reviews" as a form of "opinion spam".

Deceptive reviews, which Dixit et al. [4] also called opinion spam, have been categorized into three groups: (1) Untruthful reviews, which deliberately mislead readers or opinion mining systems; (2) Reviews on brands, which contain comments that are only concerned with the brand, or the seller of the product, and fail to review the product itself; and (3) Non-reviews, which present

either unrelated text or advertisements. The present work is mainly concerned with untruthful reviews, which are challenging for humans to detect according to Ott et al. [10]. These authors also demonstrated that distinguishing between real reviews and spam is a harder task for man than for computers.

Although the English language relies on some annotated datasets; that is, actual corpora consisting of deceptive reviews and real ones (e.g., Yelp [8], OpSpam [10]), that are useful to test algorithms for deceptive opinion categorization, the Portuguese language lacks similar datasets and methods associated with it. Looking at this opportunity, our primary aim is to assemble a dataset of deceptive book reviews in the Portuguese language for use in deceptive opinion spam detection.

There are many ways to construct an annotated corpus of deceptive texts, as studied by Gokhman et al. [5]. They divided corpora construction approaches into: a) Traditional approach, b) Non-gold standard, and c) Crowdsourcing approach. They also commented about the advantages and limitations of each approach and discussed about methods to handle plagiarism. The corpus building process used in the present work follows some steps proposed by Gokhman.

Due to the lack of a dataset of deceptive review or opinion spam for the Portuguese language, our proposal for a Master dissertation project is to create an annotated parallel corpus consisting of deceptive book reviews and real ones, both written in Brazilian Portuguese, and to use machine learning methods to detect deceptive texts correctly. Hereafter, the paper is organized as follows: Section 2 presents state-of-the-art technology to detect opinion spam. Section 3 the dataset and the developed methodology is described. Section 4 depicts the methods used herein. Section 5 presents the results and discusses them.

2 Related work

Annotated opinion review corpora are essential to measure how methods perform during opinion spam detection. The English language relies on several corpora for this purpose, but no corpus of this sort exists for the Portuguese language. Gokhman et al. [5] described two traditional strategies to create a deceptive text corpus: a) a sanctioned corpus, in which the experimenter instructs individuals to lie, or not to lie, in their reviews, and; b) an unsanctioned corpus, in which the participant lies on his or her own accord.

Ott et al. [10] created the Opinion Spam dataset, called OpSpam, which is the first gold standard labeled dataset for opinion spam. To create this dataset, deceptive opinions were obtained through the Amazon Mechanical Turk crowdsourcing service, whereas the truthful opinions were mined from TripAdvisor. In their paper, Ott et al. [10] showed that it is hard for humans to detect deceptive opinion spam, but when these authors applied Support Vector Machines (SVM) and Linguistic Inquiry and Word Count (LIWC)+bigram, they achieved 89.8% accuracy. Later, Hauch et al. [6] analyzed several studies about deceptive text identification. Their work compared several linguistic cues, most of which were psycholinguistic features that can be analyzed with the aid of the LIWC

tool [12]. This research showed some linguistic cues that help to differentiate between liars and truth-tellers.

In a recent paper, Castañeda et al. [7] showed the performance of the Latent Dirichlet Allocation (LDA) with word-space model, which outperformed almost all the other models in terms of deceptive text detection. Castañeda et al. obtained an average accuracy of 86% in one-domain setting, 75% in mixed-domain setting, and 52 to 64% in the cross-domain setting.

3 Dataset: assembling a deceptive book review corpus

While truthful opinions are ubiquitous online, deceptive opinions are difficult to obtain without resorting to heuristic methods [10]. In this section, we describe how we assembled a deceptive book review dataset.

Deceptive reviews: We assembled this dataset by using the sanctioned approach of Gokhman et al. [5] , which proposes requesting deceptive book reviews. We explicitly asked the participants to lie about a book while writing a review about it. We assembled a deceptive corpus by using Google Forms[1].We used this service to collect deceptive opinion spam at no cost. We asked the participants to write a book review consisting of 300 characters or more about any of their preferred books and to disclose their age, occupation, and gender. These features might be useful for future research.

Up to this moment, we have gathered 68 deceptive reviews. We have measured the length of the reviews and their average number of words ($\overline{w}$) and characters ($\bar{l}$) is $\overline{w} = 108.94$ words and $\bar{l} = 636.76$ characters, respectively.

Truthful reviews: We manually mined these reviews from Skoob[2], a four-million-user Brazilian book-oriented social network. For each deceptive review, we retrieved a truthful one about the same book; the pairs of truthful and deceptive reviews showed the same sentiment polarity and had a similar number of words.

We gathered 68 truthful reviews with $\bar{l} = 648.45$ characters and $\overline{w} = 114.23$ words, on average.

4 Automated approaches for review spam detection

We tested two basic automated approaches for deceptive review detection, both of which were based on a trained classifier, SVM, and Random Forest. We also used the ten-fold cross-validation approach as a validation process. The features supplied to these classifiers are shown below.

Linguistic Inquiry and Word Count (LIWC) and Brazilian LIWC

Linguistic Inquiry and Word Count (LIWC) [11] is a word-counting-based tool that relies on manually labeled word groups. LIWC classifies words into emotional, cognitive, and structural component categories. Although LIWC was not

[1] Link to the form: https://goo.gl/MfuwRB
[2] https://www.skoob.com.br/

specifically designed to assess deception, Newman et al. [9] and Hauch et al. [6] showed that its psycholinguistic features may be used to detect deceptive texts.

In 2013, Balage et al. [1] translated the English lexical LIWC into Portuguese. They also evaluated it for a Brazilian Portuguese sentiment classification task. Since Newman and Hauch used original LIWC to detect deceptive texts in English, we will use the Brazilian Portuguese LIWC to detect deceptive reviews.

Text categorization

Following Ott's nomenclature [10], we use frequency n-gram-based features for the text categorization approach. These features allow us to model the context of both the content and the word. In this work, we employ the bag-of-words (unigram) approach as feature to analyze lowercased and unstemmed texts.

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) [2] is a probabilistic generative model for discrete data collections, such as text collection, that represents documents as a mix of topics. In this case, a topic is defined as a set of words that often occur in semantically related documents. As a result, each document can combine different topics. In contrast to LIWC, which is a manually annotated lexicon, LDA generates the topics automatically, so it does not depend on language.

Support Vector Machine

One of the classifiers used here is the support vector machine (SVM), which has been continuously used for text classification [8,10]. The SVM algorithm finds a high-dimensional separating hyperplane between two data groups. As a simple example, consider a linear SVM, as described in Equation 1, that can be used as a classifier, where $\boldsymbol{w}$ is the weight vector, b is the bias learned, and $\boldsymbol{x}$ is the feature vector.

$$\hat{y} = \operatorname{sgn}(\boldsymbol{w} \cdot \boldsymbol{x} + b) \qquad \operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \end{cases} \tag{1}$$

Random Forest

Random forest is an ensemble machine learning algorithm; more specifically, it is a combination of tree predictors, so that each tree depends on the values of an independently sampled random vector with the same distribution for all trees in the forest [3]. In other words, random forest is a set of decision trees that are used to determine the final outcome.

5 Results and Discussion

LIWC, bag-of-words (BOW), and LDA have been used to construct the features used by both classifiers, SVM and Random forests. We evaluated the performance of both classifiers by stratified 10-fold cross-validation, which divides the corpus into ten folds while maintaining the same proportion of deceptive and truthful texts. We used nine of the folds to train the classifiers and the remaining fold to test it.

The LIWC features were formed from a list of all its n topics $t_0, t_1, ..., t_n$. Consider R as the set of all reviews. For each review r_i, deceptive or not, we used the proportion of words of each topic t_n was used to assemble an LIWC vector $v_{r_i} = (\alpha_1 t_1, \alpha_2 t_2, \ldots, \alpha_n t_n)$, where α_n is the proportion of words in a topic t_n in the review r_i.

To generate LDA features, we first trained an LDA model by using all the reviews r_i containing between 2 and 50 topics (m) as parameters. Then, for each trained model, we created a list containing all the m topics $t_0, t_1, ..., t_m$ trained. For each review r_i we assembled an LDA vector $v_{r_i} = (\beta_1 t_1, \beta_2 t_2, ..., \beta_m t_m)$ where $\beta_m = 1$ if the topic t_m appeared in the review r_i, or $\beta_m = 0$. The SVM and the Random forest classifier were applied to all $m = [2, 50]$ topics. The best results for LDA were obtained for $m = 9$ topics.

Table 1. Automated classifier performance for SVM and Random Forest based on stratified 10-fold cross-validation experiments, using F1-score over deceptive reviews.

Features	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
BOW + SVM	**66.7**	73.5	61.9	65.8
BOW + Random Forest	65.1	**76.3**	55.7	59.9
LIWC + SVM	52.0	51.8	**92.9**	66.2
LIWC + Random Forest	61.4	64.9	54.1	57.9
LDA + SVM	62.3	61.2	75.2	66.7
LDA + Random Forest	62.7	62.0	78.3	**67.8**

Table 1 lists some results. The best overall accuracy, 66.7%, was obtained by considering the vector of features formed from the BOW+SVM. If we avoid false positives, the BOW + Random Forest approach gets almost the same accuracy, 65.1%, but with higher precision 76.3%. Although these results seem discouraging if compared to others; e.g., 89.8% [10] and 91.2% [7], we suspect our corpus still has a small number of reviews, which came from many different books. We intend to collect more deceptive and truthful reviews and to continue to research other features and approaches to detect deceptive reviews.

6 Acknowledgement

This research was supported by the São Paulo State Research Foundation (FAPESP).

References

1. Balage Filho, P.P.P., Aluísio, S., Pardo, T.T.: An Evaluation of the Brazilian Portuguese LIWC Dictionary for Sentiment Analysis. 9th Brazilian Symposium in Information and Human Language Technology – STIL pp. 215–219 (2013)
2. Blei, D.M., Edu, B.B., Ng, A.Y., Edu, A.S., Jordan, M.I., Edu, J.B.: Latent Dirichlet Allocation. Journal of Machine Learning Research **3**, 993–1022 (2000)
3. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
4. Dixit, S., Agrawal, A.J.: Survey on review spam detection. International Journal of Computer & Communication Technology **4**(2), 68–72 (2013)
5. Gokhman, S., Hancock, J., Prabhu, P., Ott, M., Cardie, C.: In Search of a Gold Standard in Studies of Deception. Computational Linguistics pp. 23–30 (2012)
6. Hauch, V., Blandón-Gitlin, I., Masip, J., Sporer, S.L.: Are Computers Effective Lie Detectors? A Meta-Analysis of Linguistic Cues to Deception. Personality and Social Psychology Review **19**(4), 307–342 (2015)
7. Hernández-Castañeda, Á., Calvo, H., Gelbukh, A., Flores, J.J.G.: Cross-domain deception detection using support vector networks. Soft Computing **21**(3), 585–595 (2017)
8. Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.: What Yelp Fake Review Filter Might Be Doing? Seventh International AAAI Conference on Weblogs and Social Media pp. 409–418 (2013)
9. Newman, M.L., Pennebaker, J.W., Berry, D.S., Richards, J.M.: Personality and Social Psychology Bulletin. Society for Personality and Social Psychology, Inc. **29**(5), 665–675 (2003)
10. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: Finding Deceptive Opinion Spam by Any Stretch of the Imagination. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics p. 11 (2011)
11. Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., Booth, R.J.: The Development and Psychometric Properties of LIWC2007 The University of Texas at Austin. Development **1**(2), 1–22 (2007)
12. Tausczik, Y.R., Pennebaker, J.W.: The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology **29**(1), 24–54 (2010)