Automatic identification of false opinions in social networks

Identificación automática de opiniones falsas en redes sociales

GUZMAN-CABRERA, Rafael, HERNÁNDEZ-RAYAS, Angelica, PRASAD-MUKHOPADHYAY, Tirtha and RUIZ-PINALES, José

ID 1st Author: *Rafael, Guzman-Cabrera* / **ORC ID:** 0000-0002-9320-7021

ID 1st Co-author: *Angelica, Hernández-Rayas* / **ORC ID:** 0000-0002-2844-2922

ID 2nd Co-author: *Tirtha, Prasad-Mukhopadhyay* / **ORC ID:** 0000-0002-2707-390X

ID 3rd Co-author: *José, Ruiz-Pinales* / **ORC ID:** 0000-0003-2639-1487

DOI: 10.35429/EJRS.2023.17.9.36.42 Received July 30, 2023; Accepted December 30, 2023

Abstract

This paper presents the problem of detecting false opinions in social networks, also called "opinion spam", describing how lies can be automatically detected using different methods. It has been shown that deception is frequently present in everyday communication in social networks. Deception detection is a well-known challenging problem in any research area, basically because the human ability to detect deception is deficient. Particular studies on social psychology and communications show that the accuracy rates of people's abilities to detect deception are in the range of 55 to 58%, i.e., slightly better than chance. This paper addresses the specific problem of deception detection in communication. Emphasis is placed on those approaches that use affective resources such as categorical and psychometric information provided by natural language processing tools. Finally, we focus on the identification of opinion spam, whose detection is very important for reliable opinion mining. Results obtained using different machine learning methods are presented. The results obtained allow us to see the feasibility of the proposed methodology to carry out the detection of false opinions in social networks by obtaining accuracy values higher than 80%.

Options, Social Networking, Deception

Resumen

En este trabajo se presenta el problema de la detección de opiniones falsas en redes sociales, también llamadas "opinión spam", describiendo cómo las mentiras pueden detectarse automáticamente usando diferentes métodos. Se ha demostrado que el engaño está frecuentemente presente en la comunicación cotidiana en redes sociales. La detección de engaños es un problema desafiante bien conocido en cualquier área de investigación, básicamente porque la capacidad humana para detectar engaños es deficiente. Estudios particulares sobre psicología social y comunicaciones muestran que las tasas de precisión de las habilidades de las personas para detectar el engaño están en el rango de 55 a 58%, es decir, ligeramente mejor que el azar. En este trabajo se aborda el problema específico de la detección del engaño en la comunicación. Se hace hincapié en aquellas aproximaciones que utilizan recursos afectivos como la información categórica y psicométrica proporcionada por las herramientas de procesamiento del lenguaje natural. Finalmente, nos centramos en la identificación de opinión spam, cuya detección es muy importante para una minería de opinión fiable. Se presentan resultados obtenidos utilizando distintos métodos de aprendizaje automático. Los resultados obtenidos permiten ver la viabilidad de la metodología propuesta para llevar a cabo la detección de opiniones falsas en redes sociales al obtener valores de precisión superiores al 80%.

Opciones, Redes sociales, Engaño

Citation: GUZMAN-CABRERA, Rafael, HERNÁNDEZ-RAYAS, Angelica, PRASAD-MUKHOPADHYAY, Tirtha and RUIZ-PINALES, José. Automatic identification of false opinions in social networks. ECORFAN Journal-Republic of El Salvador. 2023. 9-17:36-42.

* Correspondence to Author (Email: guzmanc@ugto.mx)
† Researcher contributing first author.

Introduction

Opinion is a natural act of human beings and allows them to discern the reality that surrounds them and then take action on it. The fact that people are receiving false information is not something new and exclusive to our era, however, it has become popular due to the use of forums, blogs and social networks in general. With the general use of information technologies, it is increasingly common for users to write their opinions for or against the products or services they have purchased. These references commonly written on social networks are helpful to other consumers who wish to purchase some similar products or services. They also help manufacturers or service providers to identify new areas of opportunity on the part of consumers and allow them to know not only the opinion about them, but also to see their uses, habits, and satisfaction, among others. Consumer reviews are used by consumers to receive information about products, such as quality and usefulness, and are also used to provide data about their own experience with the product to other consumers.

In today's age of digital communications it is possible to purchase almost any product and contract all kinds of services without ever having to cross a single word with anyone. The problem of opinion detection in unstructured texts is to detect opinions that do not follow an established structure or format. This can be clearly observed in the opinions that are given on social networks such as Facebook, Instagram, Twitter, etc. Another clear example can be seen in the reviews that people give when buying a product in online shops, showing their satisfaction or dissatisfaction with the product or item they have purchased. Consequently, in order to be able to organise and filter all this type of information, new tools are needed to enable us to make the best decision regarding the purchase or rejection of these products or services. All this leads to the big problem of fake reviews (opinion spam), which are deliberately written to promote or discredit a product or service. These are reviews written by people who have not purchased a product or service, but were hired to write misleading reviews [1]. The consequences of fraudulent reviews in e-commerce range from loss of reputation and sales and apply to both product or service providers operating in the traditional way with established businesses as well as those operating online.

The challenge of this task lies in the fact that it is complicated to carry out this detection, as users express their opinions in a subjective way, in addition to the fact that each person's criteria can vary significantly, some being more direct and explicit, and others the opposite, falling into ambiguity and expressing themselves in an indirect way.

There are different techniques that can help to solve this problem, such as natural language processing and machine learning. These techniques include word tokenisation, emotion detection, text classification, among others. It is also possible to apply approaches in deep learning models, such as recurrent neural networks, in order to extract contextual information and improve the accuracy of the identification of this type of text.

This paper presents results from four experiments using the Deceptive Opinion Spam corpus, which consists of 1,600 opinions in total, the opinions are about hotel service, divided into two main categories: truthful opinions and deceptive opinions. Each category has 800 documents. Each of the four experiments is described in the methodology section.

Detecting fake reviews in hotels has become an increasingly relevant challenge. With the rise of online review platforms such as TripAdvisor or Booking, and more recently Airbnb, travellers rely heavily on the opinions of other users to make informed decisions about where to stay. However, this ease of access to information has also given rise to a growing problem: fake reviews. Fake hotel reviews are misleading or manipulated reviews that seek to distort the image of an establishment or promote hidden interests. They can come either from unfair competitors seeking to damage a hotel's reputation, or from companies hired specifically to create fake positive reviews in order to increase their ranking and attract more customers.

A study published by [2] focused on the analysis of linguistic and structural features of fake hotel reviews, using machine learning techniques to extract features such as the length of reviews, the frequency of use of certain words and the consistency of sentence structure. Through the application of these models, they were able to accurately identify a high percentage of fake reviews.

Another approach can be found in [3], where the authors proposed a method based on the analysis of the temporal evolution of reviews and the detection of "suspicious" patterns. By observing the distribution of ratings and sudden changes in opinions over time, they were able to identify patterns that indicated the presence of false opinions.

To this point, we can see the relevance of carrying out the identification of false opinions issued by users in social networks, some related work is presented below.

Related work

In [4] the authors propose a deep learning approach to detect and classify misleading opinions in online reviews. The approach involves preprocessing techniques, word representations and various machine learning models, including Naive Bayes, Logistic Regression, Support Vector Machine, Stochastic Gradient Descent and deep neural networks such as Convolutional Neural Networks (CNN), Short-Term Memory Model (LSTM), Bidirectional LSTM, Recurrent CNN and Bidirectional LSTM with Attention. The proposed approach is compared with other text classification methods and state-of-the-art approaches, and the results show that Bidirectional LSTM with Attention outperforms the other approaches.

According to the work of [4], the Attention-based Bidirectional model is considered better compared to other deep learning models due to its ability to capture the most important semantic information in the text sequence. In addition, the model uses a bidirectional neural network that retains contextual information in both directions and an attention layer that extracts only the important word representations needed to understand the meaning of the sentence.

A possible limitation of the method proposed in this paper is that it is based on a specific dataset and may not be generalisable to other datasets or domains. In addition, the method may require a large amount of labelled data to train deep learning models, which may be costly and difficult to obtain in some cases.

In [5] a reliable recommendation framework is proposed using the content features of the Deceptive opinion spam corpus dataset by using several deep learning algorithms to predict the veracity of reviews. The proposed hybrid CNN-LSTM combination involving content features. The main challenge of a recommender system lies in the reliability of the user's choices and needs.

The methodology of this paper is based on the analysis of content features, such as review text and composite score, to predict the trustworthiness of reviews. The methodology proposed by the authors focuses on improving the trustworthiness and stability of the recommender system by avoiding misleading reviews.

An interesting approach can be found in [6], in this work the authors propose a methodology based on the PU (Positive Unlabeled) learning approach which stands out for being a type of learning with positive labels and unlabeled data, this learning method is used in this work to detect misleading opinions in online reviews. This approach uses a small set of examples of misleading opinions and a set of unlabelled opinions to build accurate classifiers. The proposed method is a two-step iterative process in which a classifier is trained using a set of positive examples and a set of unlabelled data, and then this classifier is used to classify the unlabelled data set. The process is repeated until a stopping criterion is reached and the last classifier constructed is returned as the final classifier. Later, this work was modified in [7] where the authors propose the use of n-charactergrams as features for false opinion detection. They perform two experiments to evaluate the effectiveness of this approach. In the first experiment, they compare the performance of character n-grams with word n-grams in detecting misleading opinions. In the second experiment, they evaluate the robustness of the character n-gram approach when only a few examples of deceptive opinions are available for training. The authors use the Naive Bayes classifier to evaluate the performance of the proposed approach. Furthermore, they compare their approach with other existing approaches, such as sentiment analysis and spam detection.

In the study by [8], the researchers proposed an unsupervised approach for detecting false and misleading opinions using the Deceptive Opinion Spam corpus. Their goal was to identify textual patterns inherent in opinions that would allow distinguishing between genuine and misleading opinions without the need for a labelled training dataset.

The researchers explored multiple linguistic and structural features of the opinions in the corpus, such as the use of emotional words, the length of reviews, the amount of punctuation and the frequency of specific words. They then used machine learning algorithms, such as SVM (Support Vector Machines) and Naïve Bayes, to classify the reviews as genuine or false. The following section describes the methodology implemented in this work.

Methodology

It is clear that the issue of identifying false opinions is still an open research topic, as shown in the previous section, this problem is not new and this has allowed different research groups to make their contributions on different approaches that contribute to the solution of this problem. In this paper we present a methodology, see figure 1, with which competitive results are obtained when carrying out the identification of false opinions. It is worth mentioning that this methodology can be used with other corpora. A brief description of the corpus used in the experimental part is presented below.

The corpus used in the present work is: "Deceptive Opinion Spam" which is available (https://www.kaggle.com/datasets/rtatman/deceptive-opinion-spam-corpus) and consists of a dataset used in the research on the detection of false and misleading opinions. This corpus consists of hotel reviews written by real users, but with a distinction between truthful and deceptive reviews, specifically it contains: 400 truthful positive reviews from TripAdvisor [9], 400 deceptive positive reviews from Mechanical Turk [9], 400 truthful negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp [3] and 400 deceptive negative reviews from Mechanical Turk [3]. In total there are 1,600 reviews, divided into two main categories: "truthful" reviews and "deceptive" reviews. Each category has 800 documents. Illustrative examples of these opinions are:

- Truthful: "I recently stayed at this hotel during my business trip and I must say it exceeded my expectations. The staff was friendly and accommodating, the room was clean and comfortable, and the location was convenient. I highly recommend this hotel for both business and leisure travelers."

- Deceptive: "I had the worst experience at this hotel. The staff was rude and unhelpful, the room was dirty and uncomfortable, and the location was terrible. I would never recommend this hotel to anyone. Stay away!"

As you can see it is not easy to identify at first glance the false opinion, but if we look more closely we see that the negative evaluations that are made are very general and that, when a person complains about something, they usually specify in more detail what they did not like about the room.

These opinions are the ones that feed the first block of the proposed methodology, shown in figure 1.

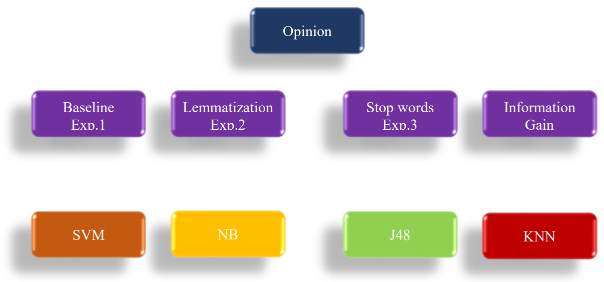

Figure 1 Methodology implemented for the identification of false opinions

Four experiments were carried out, which are described below:

- Exp 1: Baseline, the data set is taken without preprocessing.

- Exp 2: Lemmatisation of the dataset is carried out

- Exp 3: Stopwords and words with frequency less than 3 are eliminated.

- Exp 4: Information gain is applied to the set of Exp 3.

It is worth mentioning that this process does not have to be sequential, however, as will be shown in the results section, a gradual improvement is observed when moving from one to the other. Four different learning methods were used, which are described below:

- SVM: SMO divides the optimisation problem into smaller subproblems and solves them sequentially to find the hyperplane that best separates the different classes in the dataset. It is able to handle both binary and multi-class classification problems. [10, 11].

- NB: Naive Bayes is a probability-based statistical learning model, which has as its main foundation that all attributes are completely independent given the class of study [12]. Although this assumption is not regularly respected in many real-world applications, Naive Bayes remains one of the best classification algorithms today due to its simplicity and efficiency. Given a test instance d, represented by a vector of attributes (w1, w2, ..., wm), the probabilistic condition P(d|c) is computed as follows:

$$p(d|c) = \prod_{i=1}^{m} P(w_i|c)$$

- J48: is a widely used algorithm in machine learning, which belongs to the family of decision tree algorithms. This algorithm, a variant of ID3, differs in its ability to accept continuous and categorical attributes when constructing the decision tree [13]. In order to reduce classification error caused by high noise or detailed data sets, the J48 algorithm uses an improved tree pruning technique. In addition, this algorithm employs a greedy divide-and-conquer approach to recursively induce decision trees containing the attributes of the database or dataset for further classification [13]. The algorithm shows the ability to accept both continuous and categorical attributes during the construction of the decision tree and can be developed using a top-down or bottom-up approach. In addition, the algorithm splits a dataset based on the different attribute values present in the data to separate out a likely prediction.

- KNN: refers to the k-nearest neighbour classification algorithm. It is a supervised learning algorithm that is used to classify new data points based on their similarity to their nearest neighbours in the training set [14]. The algorithm uses a training data set containing examples with their respective class labels. When presented with a new data point to classify, the algorithm searches for the k nearest neighbours in the training set and assigns the new point the most frequent class among those neighbours. The value of k determines the number of nearest neighbours to be used for classification. Once the nearest neighbours are found, some distance metric can be used to calculate the similarity between the new point and the neighbours.

According to [15] there are common ways to evaluate the results of machine learning experiments, among these metrics is accuracy. Precision is a metric used in classification problems to measure the proportion of correctly identified positive cases among all the cases classified as positive by the model:

$$P = \frac{TP}{TP + TF}$$

Where: TP (True Positive) is the number of positive cases that have been correctly identified and FP (False Positive) is the number of negative cases that have been incorrectly classified as positive.

Results

The results obtained for the four experiments are shown in table 1 and figure 2.

	SVM	NB	J48	KNN
Exp.1	79.63%	12.25%	54.16%	35.32%
Exp.2	80.20%	13.43%	55.46%	36.84%
Exp.3	81.30%	77.23%	56.53%	37.78%
Exp.4	82.32%	78.50%	55.35%	38.62%

Table 1 Results obtained: accuracy metrics in the deceptive opinion spam corpus

As can be seen, the best results are obtained for SVM. The dimension of the feature vector for each experiment was as follows: for the baseline the feature vector consisted of a total of 9604 elements, in the case of experiment 2 where the lemmatisation was carried out the dimension of the vector was 9604 elements, in the case of experiment 3, where the empty words or stopwords were eliminated as well as words with a frequency of less than 3, the dimension of the vector was reduced to 3217 elements and finally in experiment 4, when using information gain, there were only 1816 elements (considering only those elements with an information gain greater than zero), it is worth noting that although the dimension of the feature vector is significantly better (it contains only 19. 21% of the instances that were in the baseline) the accuracy value is significantly increased. It also highlights the values obtained with Naive Bayes, which is observed that with a large feature vector (experiments 1 and 2) gives very bad results and improves significantly when the feature vector reduces its dimension.

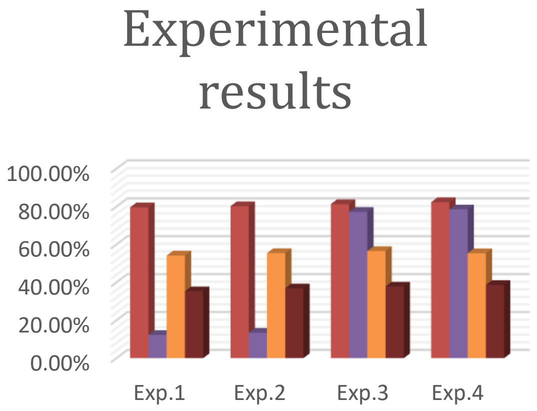

Figure 2 Results obtained: accuracy metrics in the deceptive opinion spam corpus

Conclusions

The methodology implemented in this paper addresses the detection of false opinions in the hotel industry using the Deceptive Opinion Spam corpus. Four different machine learning methods were used and results are presented using several preprocessing methods, including word lemmatisation. The best results in the evaluation metrics used are obtained with SVM in the detection of false and misleading opinions. This level of accuracy is competitive and suggests that the approach used in this study has the potential to detect misleading opinions effectively.

Importantly, word lemmatisation applied in data preprocessing has been shown to be an effective technique for improving the accuracy of false and misleading opinion detection. By reducing words to their base form, a more generalised representation is achieved and the essential features of opinions are better captured. It is important to mention a significant difference between the approach used in this paper and the approach of the authors mentioned in the literature. While the authors focused only on the detection of two classes of opinions, genuine and misleading, in this work the detection of opinions is performed in four different classes False Positive Opinions (FP), True Positive Opinions (TP), False Negative Opinions (FN), True Negative Opinions (TN). This adds complexity to the problem, as it involves classifying opinions into more categories, which is more challenging.

References

1. Fitzpatrick, E., J.C. Bachenko, and T. Fornaciari, *Automatic detection of verbal deception.* 2015.

2. Ren, Y. and D. Ji, *Learning to detect deceptive opinion spam: A survey.* IEEE Access, 2019. **7**: p. 42934-42945.

3. Ott, M., C. Cardie, and J.T. Hancock. *Negative deceptive opinion spam*. in *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies*. 2013.

4. Salunkhe, A., *Attention-based Bidirectional LSTM for Deceptive Opinion Spam Classification.* arXiv preprint arXiv:2112.14789, 2021.

5. SujithraKanmani, R. and B. Surendiran, *Boosting credibility of a Recommender System using Deep Learning Techniques-An Empirical Study.*

6. Fusilier, D.H., et al. *Using PU-learning to detect deceptive opinion spam*. in *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*. 2013.

7. Fusilier, D.H., et al. *Detection of opinion spam with character n-grams*. in *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II 16*. 2015. Springer.

8. Asghar, M.Z., et al., *Opinion spam detection framework using hybrid classification scheme.* Soft computing, 2020. **24**: p. 3475-3498.

9. Ott, M., et al., *Finding deceptive opinion spam by any stretch of the imagination.* arXiv preprint arXiv:1107.4557, 2011.

10. Tanveer, M., et al., *Comprehensive review on twin support vector machines.* Annals of Operations Research, 2022: p. 1-46.

11. Chinguel Tineo, S.F., *Evaluación de rendimiento de algoritmos en la identificación de ataques a sitios web utilizando logs de servidor.* 2022.

12. Wang, S., L. Jiang, and C. Li, *Adapting naive Bayes tree for text classification.* Knowledge and Information Systems, 2015. 44: p. 77-89.

13. Shadi Aljawarneh, M.B.Y.y.M.A., *An enhanced J48 classification algorithm for the anomaly.* Cluster Computing, 2017: p. 10549–10565.

14. Bishop, C.M., *Pattern Recognition and Machine Learning*. 2006, Springer: New York. p. 738.

15. Powers, D.M.W., *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.* International Journal of Machine Learning Technology, 2011: p. 37-63.

.