International Journal of Research in Engineering, Science and Management
Volume-1, Issue-12, December-2018
www.ijresm.com | ISSN (Online): 2581-5792

Fake Review Detection System Using Machine Learning

Aishwarya M. Kashid[1], Ankita K. Lalwani[2], Samiksha S. Gaikwad[3], Rajal A. Patil[4], R. G. Sonkamble[5], S. S. More[6]

[1,2,3,4]*B.E. Student, Department of CSE, Sanjay Ghodawat Institute, Atigre, India*
[5,6]*Professor, Department of CSE, Sanjay Ghodawat Institute, Atigre, India*

***Abstract*: In the last few years review sites are more and more confronted to spread of misinformation, to promote or to damage certain businesses various opinion spam's are done either to mislead the human readers or the sentiment analysis or opinion mining systems which are automated. In the last few years because of this reason various approaches have been proposed so that the credibility of the user generated content can be assessed. The analysis of the main review and the reviewer-centric features are proposed to detect the fake reviews by using supervised machine learning approaches rather than the unsupervised approaches which are based on graphical methods.**

***Keywords*: Fake Review, Spam, Unsupervised method, supervised method, Review Centric Features, Reviewer Centric Features, Labeled data, Classifier.**

1. Introduction

The user generated content is increasing popularity on the social websites without any form of trusted external control and thus there are no means to verify which content generated by the user is believable or which source is reliable. The consequences of spread of such misinformation are negative and it causes harm to user as well as businesses. The different subset of characteristics i.e. features often considered by various approaches connected to both reviews and reviewers as well as to the network structure linking distinct entities on the review-site in exam. The main purpose is to provide analysis of the main review and review -centric features that have been proposed to detect fake reviews, in particular approaches that employ supervised machine learning techniques. Fake reviews, fake comments, fake blogs, fake social networking postings, deceptive messages are identified by opinion spam detection. The review-centric sites such as yelp can be considered while detecting fake review detection. Unsupervised approaches have been incorporated so far for detecting fake reviews which are based on graphical methods but are not much reliable. The supervised techniques consider distinct features generated from the reviews as well as the behavior of the reviewer. A publically available large Scale and generated dataset has been considered provided by the yelp reviews which are classified using few well known supervised classifiers which bifurcate the reviews as true or deceptive by considering various features of the data.

2. Literature survey

In the Social Web, evaluating information credibility deals with the analysis of the user-generated content (UGC), the authors' characteristics, and the intrinsic nature of social media platforms, i.e., the social relationships connecting the involved entities. These characteristics, namely features, can be simple linguistic features associated with the text of the UGC, they can be additional meta-data features associated for example with the content of a review or a tweet, they can also be extracted from the behavior of the users in social media, i.e., behavioral features, or they can be connected to the user profile (if available). Furthermore, different approaches have taken into consideration product-based features, in the case of review sites where products and/or services are reviewed, or have considered social features, which exploit the network structure and the relationships connecting entities in social media platforms. In the last years, several approaches have been proposed to assess in an automatic or semi-automatic way the credibility of information in the Social Web; in particular, the most investigated tasks have been the identification of:

- Opinion spam in review sites.
- Fake news in micro blogging sites.
- Potentially harmful/inaccurate online health information.

By considering the effectiveness of supervised solutions, discussions and analysis on a general level the most appropriate review- and reviewer-centric features that have been proposed so far in the literature to detect fake reviews moreover, it proposes some new features suitable for this aim, in particular to detect singleton fake reviews. To avoid the problem of the limited size of the labeled datasets considered up to now by the literature, large-scale publicly-available datasets have been employed for evaluation purposes.

3. Proposed methodology

We provide a global overview of the various features that can be employed to detect fake reviews. Since the most effective approaches in the literature are in general supervised and consider review- and reviewer-centric features, these two classes will be taken into consideration.

A. Review-centric Features

The first class of features that have been considered is constituted by those related to a review. They can be extracted both from the text constituting the review, i.e., textual features, and from meta-data connected to a review, i.e., meta-data features. A large part of reviews are singletons, i.e., there is only one review written by a given reviewer in a certain period of time for this kind of reviews, specific features must be designed.

1) Textual Features

It is possible to use Natural Language Processing techniques to extract simple features from the text, and to use as features some statistics and some sentiment estimations connected to the use of the words. Several approaches employ as textual features both unigrams and bigrams extracted from the text of reviews. Statistical data like

- Number of words,
- Ratio of capital letters,
- Ratio of capital words,
- Ratio of first person pronouns,
- Ratio of 'exclamation' sentences,
- A number representing the proportion of subjective words.

2) Meta-data Features:

They can be generated by reasoning on the review's cardinality with respect to the reviewer and the entity reviewed. These features include:

- Basic features like
- Rating of review,
- Rating deviation, i.e., the deviation of the evaluation provided in the review with respect to the entity's average rating
- Singleton feature
- Burst features which can be either due to sudden popularity of the entities reviewed or to spam attacks.

B. Reviewer-centric features

This group of features is composed of features related to the reviewer's behavior. In this way it is possible to go beyond the content and meta-data associated with a review, which are limited for classification, and considering the behavior of users in general in writing reviews.

1) Textual features

1) The textual features are employed to address the problem of review duplication. The following textual features have been taken
   - Maximum Content Similarity (MCS), i.e. the evaluation of the maximum similarity over the user's reviews.
   - Average Content Similarity (ACS), i.e., the evaluation of the average similarity over the user's reviews.
   - Word number average, i.e., the average number of words that the user utilizes in his/her reviews

2) Rating features

They are based on some aggregation, for each considered reviewer, of the information concerning the ratings

- Total number of reviews.
- Ratios, i.e., the ratio of negative, positive and 'extreme' reviews.
- Average deviation from entity's average.

3) Temporal features

They are based on the temporal information that further describes how the ratings are distributed over the time

- Activity time of the user the difference of timestamps of the last and first reviews for a given reviewer.
- Maximum rating per day
- Data entropy, the temporal gap in days between consecutive pairs or reviews.

The following techniques are used for implementing the supervised machine learning technique for classification, for balancing data, and for testing the classifier.

C. Choice of the classifier and implementation

The majority of supervised classifiers to tackle the issue of opinion spam detection are based on Naıve Bayes or Support Vector Machines (SVM). To implement the classifier, the Python programming language has been employed, as it is used by a large community of developers, thus offering a vast set of tools and libraries for different aims.

D. Choice of the dataset

The classification provided by Yelp has been used as a ground truth, where recommended reviews correspond to 'genuine' reviews, and not recommended reviews correspond to 'fake' ones. The strengths of these datasets are

- The high number of reviews per user, which allows to consider the behavioral features of each user
- The diversified kinds of entities reviewed, i.e., restaurants and hotels
- The datasets only contain basic information, such as the content, label, rating, and date of each review, connected to the user who generated them.

E. Balancing data

Imbalanced data represents one of the major issues that have to be tackled when performing supervised classification. In the training phase, if the unbalancing of training data is not considered, there is the risk that the classifier learns mainly from the largest class of labeled data therefore neglecting the minority class. The oversampling method is considered, it consists in augmenting the minority class to balance it with the largest one.

4. Block diagram

Our framework consists of three major modules:

- *Data Collection:* This module performs tasks related to gathering the information for this purpose we design

a web crawler which extracts all the links from the page. Once parsed, the information is stored in the MySQL database.

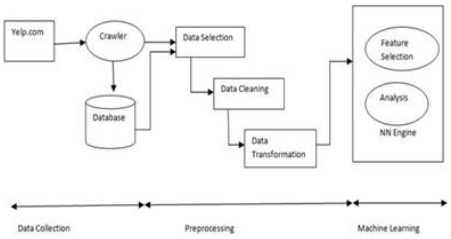

Fig. 1. Modules of the framework

- *Data pre-processing:* The data which is collected is not consistent using various machine learning algorithms the data is trained and presented in a particular format.
- *Machine Learning:* This module deals with the various feature sets under consideration and analyze them in order to obtain insights from it.

5. Conclusion

The approaches to fake review detection are based on data-driven methods that consider several features associated with reviews, reviewers, and the network structure of the social network that can be used to classify reviews in terms of their credibility. Supervised classifiers are in general more effective, and usually employ review and reviewer-centric features. Unsupervised solutions are in general less effective, but have the advantage that they do not need labeled datasets for training. Supervised solutions, on the contrary, have proven their effectiveness with respect to too small or review-site-dependent labeled datasets, and with respect to small subsets of features.

References

[1] Julian Fontanarava, Gabriella Pasi, and Macro Viviani, "Feature Analysis for Fake Review Detection through Supervised Classification," in 2017 International Conference on Data Science and Advanced Analytics. IEEE, 2017, pp. 658–666.
[2] Fangtao Li, Minlie Huang, Yi Yang and Xiaoyan Zhu "Learning to Identify Review Spam", Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, 2011.
[3] https://www.cs.uic.edu/~liub/.FBS/fake-reviews.html
[4] https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to- understand-and-implement-text-classification-in-python/