# Fake Review Detection through Supervised Classification

[1]Pankaj Chaudhary,
[1]Joint Director & Dean(Academic),
[1]JB Institute of Technology, Dehradun,

[2]Abhimanyu Tyagi
[2] Managing Director,
[2] Value Prospect Consulting, Noida[1]

[3]Santosh Mishra
[3] Associate Professor,
[3]JB Institute of Technology, Dehradun,

**Abstract—In present era, in brand selection we mainly depend on online reviews and review sites are more confronted with *the spread of misinformation, i.e., opinion spam, which aims at promoting or damaging some target businesses, by misleading either human readers, or automated opinion mining and sentiment analysis systems.***

***For this reason, in the last years, several data-driven approaches have been proposed to assess the credibility of user-generated content diffused through social media in the form of on-line reviews.***

***Distinct approaches often consider different subsets of characteristics, i.e., features, connected to both reviews and reviewers, as well as to the network structure linking distinct entities on the review-site in exam.***

***This article aims at providing an analysis of the main review and reviewer-centric features that have been proposed up to now in the literature to detect fake reviews, in particular from those approaches that employ supervised machine learning techniques. These solutions provide in general better results with respect to purely unsupervised approaches, which are often based on graph-based methods that consider relational ties in review sites.***

***Furthermore, this work proposes and evaluates some additional new features that can be suitable to classify genuine and fake reviews. For this purpose, a supervised classifier based on Random Forests have been implemented, by considering both well-known and new features, and a large-scale labeled dataset from which all these features have been extracted. The good results obtained show the effectiveness of new features to detect in particular singleton fake reviews, and in general the utility of this study.***

## I. INTRODUCTION

The social Web and the increasing popularity of social media have led to the spread of multiple kinds of content (i.e., textual, acoustic, visual) generated directly by users, the so called *user-generated content* (UGC). By means of Web 2.0 technologies, it is possible for every individual to diffuse contents on social media, almost without any form of trusted external control. This implies that there are no means to verify, a priori, the reliability of the sources and the believability of the content generated. In this context, the issue of assessing the credibility of the information diffused by means of social media platforms is receiving increasing attention from researchers.

In particular, this issue has been deeply investigated in review sites, where the spread of misinformation in the form of *opinion spam*, and the negative consequences that it brings, are particularly harmful for both businesses and users. In this context, opinion spam detection aims at identifying fake reviews, fake comments, fake blogs, fake social network postings, deceptions, and deceptive messages [1], and to make them readily recognizable. Detection techniques to identify *fake reviews* have been proposed in particular for specific review sites such as TripAdvisor1 or Yelp, 2 where users' reviews have a powerful effect on people visiting the Website for advice. Therefore, a recommendation of a product or a service such as a restaurant or a hotel based on false information can have detrimental consequences.

Most approaches that have been proposed so far to detect fake reviews in these social media platforms rely on supervised machine learning techniques and on distinct characteristics, i.e., *features*, connected to the reviews and/or the reviewers who generated them. It has been shown in the literature that their usage can lead to an effective identification of suspicious contents and/or reviewers'behaviors, and consequently of misinformation [2].

Recent approaches have suggested the additional use of features that consider the social structure of the network underlying the considered review site. These approaches, which are often based on unsupervised graph based methods, usually provide worse performance with respect to supervised solutions. On the other hand, supervised approaches too present some issues. First, available solutions have often considered a small set of features, or distinct classes of features separately; second, they have been evaluated on small datasets extracted from the well-known review sites previously cited. Thus, the proposed solutions are in most of the cases partial, or review-site-dependent.

Considering the variety of features that have been proposed and used separately by supervised approaches, the goal of this article is to provide a *feature analysis* illustrating the most suited and general *review-* and *reviewer-centric features* that can be employed in the review site context to detect fake reviews. Among these features, some are well-known and taken from the literature, others are *new* and constitute a further contribution of the paper. To evaluate the utility of this set of features in classifying genuine and fake reviews, a *supervised classifier* based on a well-known machine learning technique has been implemented. With respect to the literature, a publicly available *large-scale* and *general dataset* from the Yelp.com review site has been considered. This allows to provide more significant results with respect to the contribution of each feature taken singularly and of groups of features. In particular, the important contribution of a specific group of features in analyzing the

credibility of the so called *singleton reviews* has emerged. The promising results obtained show the effectiveness and the possible utility of the feature analysis illustrated in this article.

## II. RELATED WORK

In the last few years, depending on the context, researchers have proposed many different approaches to tackle the issue of the assessment of the *credibility* of the information diffused through social media [2]. Historically, the concept of credibility has been in turn associated with believability, trustworthiness, perceived reliability, expertise, accuracy, and with numerous other concepts or combinations of them [3].

According to Fogg and Tseng [4], credibility is a *perceived* quality of the *information receiver*, and it is composed of multiple dimensions. Different characteristics can be connected to: (*i*) the *source* of information, (*ii*) the *information* itself, i.e., its structure and its content, and (*iii*) the *media* used to disseminate information [5]. It has been demonstrated that, when considering these characteristics in terms of credibility, the impact of the delivery medium can change the perception that people have about sources of information and information itself [3], [5]. For this reason, one important question to be tackled nowadays is whether new media in the digital realm introduce new factors that may concur to credibility assessment [6], [7].

In the Social Web, evaluating information credibility deals with the analysis of the user-generated content [8], 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 [9], [10].

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: (*i*) *opinion spam* in review sites [9], (*ii*) *fake news* in microblogging sites [11], and (*iii*) potentially harmful/inaccurate *online health information* [12]. In general, the majority of these approaches focus on data-driven techniques, which classify UGC with respect to credibility by employing different models.

With regard to opinion spam detection, and in particular to fake review detection, which is the focus of this paper, the approaches that have produced the best results are generally based on supervised or semi-supervised machine learning techniques that take into account both review- and reviewercentric features. The first approaches were purely *linguistic*, in the sense that they employed simple textual features extracted from the text of reviews, often in the form of unigrams and/or bigrams [13], [14], [15], [16]. Other linguistic approaches have proposed generative classifiers based on language models [17], [18]. It has been demonstrated by Mukherjee et al. in [19] that focusing only on linguistic features is not effective to detect fake reviews from real datasets, since it is practically impossible for a human reader to distinguish between credible and not credible information by simply reading it, especially in an era where the skills in writing false reviews are constantly improving [20].

For this reason, more effective *multi-feature-based* approaches have been proposed, which employ several features of different nature in addition to simple linguistic ones, either by applying supervised or semi-supervised machine learning [1], [19], [21], or by implementing the Multi-Criteria Decision Making (MCDM) paradigm [22]. These approaches usually focus on small labeled datasets for evaluation purposes, constituted in most of the cases by 'near ground truth' data [9].

They usually avoid to consider features that are extracted from the social ties constituting the network of entities (e.g., users, products, reviews) considered by the review site. On the contrary, this kind of features is often utilized (together with the other features previously described) by graph-based approaches [23], [24]. These latter approaches are in most of the cases unsupervised, even if sometimes they can be coupled with a supervised learning phase on a limited number of classification labels [25]. With respect to supervised approaches, totally unsupervised solutions generally provide slightly worst results [2], [9], [20].

This paper, by considering the effectiveness of supervised solutions, discusses and analyzes 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, an issue that has not yet received the attention it deserves. To avoid the problem of the limited size of the labeled datasets considered up to now by the literature, two large-scale publicly-available datasets presented in [25] have been employed for evaluation purposes.

## III. FEATURE ANALYSIS

As briefly introduced in Section II, many and different are the features that have been considered so far in the review site context to identify fake reviews. In some cases, features belonging to different classes have been considered separately by distinct approaches. In other cases, the employed features constitute a subset of the entire set of features that could be taken into account; furthermore, new additional features can be proposed and analyzed to tackle open issues not yet considered, for example the detection of singleton fake reviews. For these reasons, in this section we provide a global overview of the various features that can be employed to detect fake reviews. Both significant features taken from the literature and new features proposed in this article are considered. Since the most effective approaches discussed in the literature are in general supervised and consider *review-* and *reviewer-centric features*, these two classes will be presented in the following sections. The choices behind the selection of the features belonging to the above mentioned classes will be detailed along each section. When the features are taken from the

literature, they will be directly referred to the original paper where they have been initially proposed. The absence of the reference will denote those features that have been widely used by almost every proposed technique.

Finally, the presence of the label denoted by [*new*] will indicate a feature proposed for the first time in this article. *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*. In every review site, the *time* information regarding the publication of the review, and the *rating* (within some numerical interval) about the reviewed business are metadata, are always provided. In addition, in relation to metadata features, those connected to the *cardinality* of the reviews written by a given user must be carefully studied. In fact, 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 (this means that in the user account there is only one review at the time of the analysis). For this kind of reviews, specific features must be designed. In fact, as it will be illustrated in the following, many of the features that have been proposed in the literature are based on some statistics over several reviews written by the same reviewer. In the case of singletons, these features loose their relevance in assessing credibility.

Therefore, the definition of suitable features that are effective for detecting also singleton fake reviews becomes crucial. *1) Textual Features:* as briefly illustrated in Section II, it is practically impossible to distinguish between fake and genuine reviews by only reading their content. The analysis provided by Mukherjee et al. in [19] has shown that the KL-divergence between the languages employed by spammers and non spammers in Yelp is very subtle. However, the good results obtained in [26] by using linguistic features on a domain specific dataset (i.e., a Yelp's dataset containing only New York japanese restaurants), show that at least on a domain specific level, textual features can be useful. 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.

• *Text*: several approaches employ as textual features both *unigrams* and *bigrams* extracted from the text of reviews,as illustrated in Section II.

• *Text statistics:* several statistics on the review content have been proposed as features by Li et al. in [21]:

– *Number of words*, i.e., the length of the review in terms of words;

– *Ratio of capital letters*, i.e., the number of words containing capital letters with respect to the total number of words in the review;

– *Ratio of capital words*, i.e., considering the words where all the letters are uppercase;

– *Ratio of first person pronouns*, e.g., 'I', 'mine', 'my',etc.;

– *Ratio of 'exclamation' sentences*, i.e., ending with the symbol '!'.

• *Sentiment evaluations:*

– *Subjectivity*, i.e., a number representing the proportion of subjective words (expressing sentiment, judgment) as opposed to objective (descriptive) words.

*2) Meta-data Features:* these kinds of features are extracted from the meta-data connected to reviews, or they can be generated by reasoning on the reviews' cardinality with respect to the reviewer and the entity reviewed.

• *Basic features:*

– *Rating*, i.e., the rating attributed in the review to the entity, in the form of some numerical value belonging to a given interval (e.g., 1-5 'stars');

– *Rating deviation* [27], i.e., the deviation of the evaluation provided in the review with respect to the entity's average rating;

– *Singleton* [25], i.e., it indicates the fact that the review is the only one provided by a reviewer in a given period of time (e.g., a day).

These basic features rely on some simple and intuitive heuristics. A fake review tends to contain a more 'extreme' rating with respect to genuine reviews, thus implying that the rating deviation from the entity's average rating is higher; furthermore, a singleton review provided by a user could indicate that s/he is not particularly involved in the review site community, which constitutes a possible indication of unreliability.

• *Burst features*: it is said that reviews for an entity are 'bursty' when there is a sudden concentration of reviews in a time period. These *review bursts* can be either due to sudden popularity of the entities reviewed or to spam attacks. Since it has been proven that reviews in the same

burst tend to have the same nature [28], it is possible to easily identify groups of fake reviews by analyzing the nature of the burst. Two burst detection studies have been described in [27], [28]. Taking inspiration from the just cited works, in this paper several features considering *burstiness* have been introduced. These features are related to the time window in which a review has been posted, relatively to a given reviewed entity. Basically, a review is more likely to be fake if it is posted on a day when the number of reviews is abnormally high, and when the average rating associated with an entity in a review (in a specific time window) varies significantly with respect to the entity's average rating (in general it decreases, for example passing from 3.5/5 to 2/5).

This assumption is valid for every kind of review, but it has proved to be particularly effective in boosting the detection of singleton fake reviews, which is particularly difficult without considering burstiness. With respect to the literature, the following new features have been considered:

– *Density* [*new*], i.e., the number of reviews for a given entity on a given day of publication;

– *Mean rating deviation* [*new*], i.e., the deviation of the average rating of an entity on the considered day with respect to the entity's average rating (in general);

– *Deviation from the local mean* [*new*], i.e., the numerical evaluation if the rating assigned to a given entity in a review is close to the average rating assigned on the considered day;

– *Early time frame* [27], i.e., the amount of time it takes before the first review on a given entity is posted. In fact, often spammers review early to increase the impact of their (false) opinions on the audience.

**B. Reviewer-centric Features**

This group of features is composed of features related to the reviewers' *behavior*. 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***:* when considering reviewers' behavior, textual features are employed to address the problem of review duplication, which has been mentioned and studied in several research papers [1], [29], [30]. Specifically, the following textual features have been taken from [31]:

• *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 her/his reviews. Users who duplicate their reviews will have a really high MCS. This can be considered as a suspicious behavior. Content similarity is expressed by means of the *cosine similarity* between the bag-of-words representations of the user's reviews.

**2) Rating Features***:* they are based on some aggregation, for each considered reviewer, of the information concerning the ratings and/or the reviews s/he provided:

• *Total number of reviews*;

• *Ratios* [27], i.e., the ratio of negative, positive, and 'extreme' reviews (i.e., whose rating corresponds to the extremes of the considered rating interval);

• *Average deviation from entity's average* [31], i.e., the evaluation if a user's ratings assigned in her/his reviews are often very different from the mean of an entity's rating (far lower for instance);

• *Rating entropy* [25], i.e., the entropy of rating distribution of user's (entity's) reviews;

• *Rating variance* [*new*], i.e., the squared deviation of the rating assigned by a user with respect to the ratings mean. The variance as a rating feature has been added to further describe how the ratings for a particular user are distributed.

**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* [31], i.e., the difference of timestamps of the last and first reviews for a given reviewer;

• *Maximum rating per day* [25], i.e., the maximum rating provided by a reviewer in the considered day;

• *Date entropy* [25], i.e., the temporal gap in days between consecutive pairs or reviews ;

• *Date variance* [*new*], i.e., the squared deviation of the timestamps in which a user post her/his reviews with respect to the timestamps mean. The variance as a temporal feature has been added to further describe how the reviews for a particular user are distributed over time windows.

## IV. SUPERVISED CLASSIFICATION

The classifier has been implemented to evaluate the impact of distinct features and set of features on the classification, and the overall performance that can be achieved. This constitutes an important part of the work, as it shows the impact of several kinds of features on a large-scale dataset, whereas many papers present only a small subset of these features, and evaluate them on small or review-site-dependent datasets, making it hard to assess their relative importance.

## V. CONCLUSION

In the last years, the issue of how to assess the credibility of social media information and in particular how to identify opinion spam in review sites has received increasing attention by researchers. Usually, 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 reviewand reviewer-centric features. Unsupervised classifiers are in most of cases based on graph-based models, and focuses on the social ties underlying the review site in exam (together with other kinds of 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 among the ones that have been proposed in general in the literature.

In this article, focusing on the effectiveness of supervised classification, a feature analysis has been performed, in order to summarize the main review- and reviewer-centric features that are suited for fake review detection, and to propose new features that can be particularly useful to detect singleton reviews. To evaluate the impact of these features, a supervised classifier based on Random Forests has been developed.

To avoid the issues connected to the limited volume of available ground truths, a publicly available large-scale and general labeled dataset has been employed for evaluation purposes. The promising results obtained witness the utility of the proposed study.

## REFERENCES

[1] N. Jindal and B. Liu, **"Opinion spam and analysis," in *Proceedings of the 2008 International Conference on Web Search and Data Mining"***.ACM, 2008, pp. 219–230.

[2] M. Viviani and G. Pasi, **"Credibility in Social Media: Opinions,News, and Health Information - A Survey,"** *WIREs Dat Mining and Knowledge Discovery*, 2017. [Online]. Available:http://dx.doi.org/10.1002/widm.1209

[3] C. S. Self, **"Credibility"**, in *An Integrated Approach to Communication Theory and Research, 2nd Edition*, M. B. Salwen and D. W. Stacks, Eds. Routledge, Taylor and Francis Group, 2008, pp. 435–456. [Online]. Available: http://dx.doi.org/10.4324/9780203887011

[4] B. J. Fogg and H. Tseng, **"The elements of computer credibility,"** in *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*. ACM, 1999, pp. 80–87.

[5] M. J. Metzger, A. J. Flanagin, K. Eyal, D. R. Lemus, and R. M. McCann, **"Credibility for the 21st century: Integrating perspectives on source,message, and media credibility in the contemporary media environment,"** *Annals of the International Communication Association*, vol. 27, no. 1, pp. 293–335, 2003.

[6] M. J. Metzger and A. J. Flanagin, **"Credibility and trust of information in online environments: The use of cognitive heuristics,"** *Journal of Pragmatics*, vol. 59, Part B, no. 0, pp. 210 – 220, 2013, biases and constraints in communication: Argumentation, persuasion and manipulation. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378216613001768

[8] M.-F. Moens, J. Li, and T.-S. Chua, Eds., ***Mining User Generated Content***, ser. Social Media and Social Computing. Chapman and Hall/CRC, 2014.

[9] A. Heydari, M. ali Tavakoli, N. Salim, and Z. Heydari, **"Detection of review spam: A survey,"** *Expert Systems with Applications*, vol. 42, no. 7, pp. 3634–3642, 2015.

[10] B. Carminati, E. Ferrari, and M. Viviani, **"A multi-dimensional and event-based model for trust computation in the social web,"** in *International Conference on Social Informatics*. Springer, 2012, pp. 323–336.

[11] C. Castillo, M. Mendoza, and B. Poblete, **"Predicting information credibility in time-sensitive social media,"** *Internet Research*, vol. 23, no. 5, pp. 560–588, 2012.

[12] T. J. Ma and D. Atkin, **"User generated content and credibility evaluation of online health information: A meta analytic study,"** *Telematics and Informatics*, 2016.

[13] K.-H. Yoo and U. Gretzel**, "Comparison of deceptive and truthful travel reviews,"** *Information and communication technologies in tourism 2009*, pp. 37–47, 2009.

[14] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, **"Finding deceptive opinion spam by any stretch of the imagination,"** in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 309–319.

[15] S. Banerjee and A. Y. Chua, "**Applauses in hotel reviews: Genuine or deceptive?"** in *Science and Information Conference (SAI), 2014*. IEEE, 2014, pp. 938–942.

[16] D. H. Fusilier, M. Montes-y G´omez, P. Rosso, and R. G. Cabrera, **"Detection of opinion spam with character n-grams,"** in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2015, pp. 285–294.

[17] C. L. Lai, K. Q. Xu, R. Y. K. Lau, Y. Li, and L. Jing**, "Toward a language modeling approach for consumer review spam detection,"** in *2010 IEEE 7th International Conference on E-Business Engineering*, Nov 2010, pp.1–8.

[18] R. Y. K. Lau, S. Y. Liao, R. C.-W. Kwok, K. Xu, Y. Xia, and Y. Li, **"Text mining and probabilistic language modeling for online review spam detection,"** *ACM Trans. Manage. Inf. Syst.*, vol. 2, no. 4, pp. 25:1–25:30, Jan. 2012. [Online]. Available: http://doi.acm.org/10.1145/2070710.2070716

[19] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, **"What Yelp fake review filter might be doing?"** in *Proceedings of ICWSM*, 2013.

[20] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," *Journal of Big Data*, vol. 2, no. 1, p. 23, 2015.

[Online]. Available: http://dx.doi.org/10.1186/s40537-015-0029-9

[21] F. Li, M. Huang, Y. Yang, and X. Zhu, **"Learning to identify review spam,"** in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, no. 3, 2011, p. 2488.

[22] M. Viviani and G. Pasi, **"Quantifier guided aggregation for the veracity assessment of online reviews,"** *International Journal of Intelligent Systems*, 2016.

[23] G. Wang, S. Xie, B. Liu, and S. Y. Philip, "**Review graph based online store review spammer detection,"** in *2011 IEEE 11th International Conference on Data Mining*. IEEE, 2011, pp. 1242–1247.

[24] L. Akoglu, R. Chandy, and C. Faloutsos, **"Opinion fraud detection in online reviews by network effects,"** *ICWSM*, vol. 13, pp. 2–11, 2013.

[25] S. Rayana and L. Akoglu, **"Collective opinion spam detection: Bridging review networks and metadata,"** in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 985–994.

[26] J. Fontanarava, G. Pasi, and M. Viviani, **"An ensemble method for the credibility assessment of user-generated content,"** in *WI'17 Proceedings - 2017 IEEE/WIC/ACM International Conference on Web Intelligence*, 2017, to appear.

[27] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, and R. Ghosh, **"Spotting opinion spammers using behavioral footprints,"** in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 632–640.

[28] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, **"Exploiting burstiness in reviews for review spammer detection."** *ICWSM*, vol. 13, pp. 175–184, 2013.

[29] N. Jindal and B. Liu, **"Review spam detection,"** in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 1189–1190.

[30] A. Mukherjee, B. Liu, and N. Glance, **"Spotting fake reviewer groups in consumer reviews,"** in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 191–200.

[31] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, **"Fake review detection: Classification and analysis of real and pseudo reviews,"** UICCS- 03-2013. Technical Report, Tech. Rep., 2013.

[32] H. He and E. A. Garcia, **"Learning from imbalanced data,"** *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp.1263–1284, 2009.

[33] M. Luca and G. Zervas, **"Fake it till you make it: Reputation, competition, and yelp review fraud,"** *Management Science*, vol. 62, no. 12, pp. 3412–3427, 2016.

[34] N. V. Chawla, **"Data mining for imbalanced datasets: An overview,"** in *Data mining and knowledge discovery handbook*. Springer, 2005, pp. 853–867.

[35] D. M. Hawkins, **"The problem of overfitting,"** *Journal of chemical information and computer sciences*, vol. 44, no. 1, pp. 1–12, 2004.

[36] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, **"Smote: synthetic minority over-sampling technique,"** *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

Online Anti-Opinion Spam: Spotting Fake Reviews from the Review Sequence

[1]Pankaj Chaudhary, [2] Rajat Shahni

[1]Joint Director & Dean (Academic), [2] Software Engineer,

[1]JB Institute of Technology, Dehradun, [2] Contour Software Systems Pvt. Ltd, Noida

---

***Abstract-Detecting review spam is important for current e-commerce applications. However, the posted order of review has been neglected by the former work. In this paper, we explore the issue on fake review detection in review sequence, which is crucial for implementing online anti-opinion spam. We analyze the characteristics of fake reviews firstly.***

***Based on review contents and reviewer behaviours, six time sensitive features are proposed to highlight the fake reviews. And then, we devise supervised solutions and a threshold-based solution to spot the fake reviews as early as possible. The experimental results show that our methods can identify the fake reviews orderly with high precision and recall.***

---

## I. INTRODUCTION

More and more users prefer to post reviews on products and services for sharing their opinions and experiences in the eBusiness web sites, such as Amazon. Many potential consumers would make their purchase decisions based on these online reviews. This motivates some unscrupulous merchants to apply fake reviews on misleading the potential consumers by enhancing their reputation or diminishing the competitors'. Thus, it brings an urgent demand to detect fake reviews as early as possible for reducing their influence.

The prior works on spotting review spam can be divided into three groups roughly based on the detected targets : review spam (e.g. [1], [2], [4], [5]), review spammer (e.g. [3], [6]) and the group of spammers (e.g. [ 7]). These works have pushed the anti-opinion spam forward, but they have ignored the review orders. However, identifying the review spam according to their presence orders is very important for the target of online opinion spam detection. Therefore, we focus on the problem of identifying fake reviews from the review sequences.

Our main idea is to highlight the fake reviews in the review sequences with the time sensitive features firstly based on the review contents and the reviewer behaviors. Secondly, we devise a supervised solution and a threshold based one to detect the fake reviews in the review sequences separately.

## II. HIGHLIGHTING THE FAKE REVIEWS

In this section, we propose six features updated dynamically to highlight the review spam.

### A. Modeling the review contents

Let R = {r[l], r[2], ... , r[n]} be a review sequence, and number indicates the posted order of review, the review r[i] contains multiple information: reviewer ID r[i].u, posted time r[i].t, review content r[i].c and product ID r[i].p.

**• Personal content similarity** (F 1)

If the reviewer r[i].u posts hislher own reviews repeatedly, r [i]. c would has a relative high similarity with his/her reviews. We maintain a review centroid for each reviewer, which consists of the terms' average occurrence frequencies in the reviews posted by r[i].u. Thus, we can evaluate the personal content similarity of the detected review as follows.

$$Su = similarity(r[i].c, Cr[ij.u) \quad (1)$$

where similarity is the similarity function likes cosine, Cr[ij.u is the review content centroid of r[i].u. After review r[i] is detected, the centroid Cr[iju is updated immediately.

**• Similarity with reviews on a product** (F2)

A fake review might be the duplicate or near-duplicate of an existing one on the same product. If so, it would be closely related with the product. If there are multiple reviews between the fake review and the normal ones copied, readers always can not identify the fake review because most of them prefer to read the reviews in the first few pages.

Compared with the normal reviews, fake reviews would have higher similarity with the "review centroid" of the product. Thus, we can calculate the similarity of the detected review with those on the same product as follows.

$$Sp = similarity(r[i].c, Cr[ij.p) \quad (2)$$

where Cr[ij.p is the centroid of reviews on product r[i].p, which is similar to the Cr[ij.u' The review centroid of a product would be updated after a review is processed.

**• Similarity with reviews on other products** (F3)

It is thorny to identify whether r[i] is a duplicate or near duplicate of r[j] for all reviews on different products and j < i, when i is a large number, r[i].p i- r[j].p and r[i].u i-r [j]. u.

Firstly, it is impractical to calculate the similarity of r[i].c with each r[j].c, since the total count of review pairs would be very large. Secondly, if we apply the methods like Fl and F2, the discriminating components of centroid would tend to be 0 because of so many reviews. Thus, we propose a solution for this case based on min hashing.

Given a string collection S = {Ol,' .. , Om} and a hash function h, the minhashing value of S is defined as follows.

mh(S) = argmoEiSn (h(o)) where the hash function h is used to permutate the elements in S randomly.

For a set of hash functions generated randomly, if each corresponding min hasing value of review r[i].c is equal to that of review r[j].c, r[i].c would be a duplicate of r[j].c with relatively high probability. Therefore, min hasing can measure the probability that the r[i].c contains the repeated contents in the existing reviews. By this measurement, we can determine whether r[i].c is a duplicate/near-duplicate of the existing one or not in a probabilistic sense.

Firstly, we generate d hash functions {hI, h2,' " , hd} randomly. And then, we calculate the minhashing value for each hash function, and use these minhashing values to construct a hash signature for each review content (r[i].c), namely,

Sig(r[i].c) = H(mhl (r[i].c), mh2(r[i].c), ... , mhd(r[i].c)) where H is a message-digest algorithm, which can generate a unique signature for a set of minhasing values.

The signatures of all reviews are maintained in a set. If the signature of r[i].c already exists in this set, r[i].c would be a duplicate or near-duplicate with very high probability. Otherwise, we generate multiple signatures for each review by repeating the processes above. By this way, we can evaluate the probability of that the near-duplicate's signature is different with the source review. Thus, The more times the signatures of r[i] are the same with those of r[j].c, r[i].c would be more likely to be a near-duplicate of r[j].c.

Formally, let hil, hi2,' .. , hid ( i = 1,2,⋯ , b) denote b sets of hash functions. The Hl,H2,' " ,Hb denote b signature sets respectively. Thus, the probability of r[i].c be a duplicate or near-duplicate can be evaluated as follows.

So = L□=l e x ist(Sigi) b {I Sigi E Hi e x ist( Sigi) = 0 Sigi rt. Hi

**B. Modeling the reviewer behaviors**

Some spammers' behaviors could imply that their reviews are fake, such as a spammer posts multiple reviews in a short period of time. Thus, we can model such behaviors to highlight the fake reviews.

• **The review frequency of reviewer** (F4)

Spammers prefer to post many fake reviews in a small time window for more benefits. Thus, we can measure the probability of being a fake review by the time interval between two consecutive posted time points for one reviewer.

S _ _ Iu[pre(r[i].u),r[i] .t] uf -1 max(Iu)

where pre(r[i].u) is the latest review time of r[i].u, r[i].t is the posted time of review r[i], Iu[x, y] is the time interval between x and y, and max(Iu) is the maximum time interval among all pairs of adjacent reviews posted by r[i].u.

• **The review frequency of a product** (F5)

Multiple fake reviews occurring in a short time interval would make greater impact. If a product is c Olmnented very frequently in a small time window, it might be attacked. Of cause, this could be caused by other reasons like promotions. But we also treat it as an index of spammer's behavior.

S -1- Ip[pre(r[i].p ),r[i] .t]
pf - max(Ip)

where pre(r[i].p) is the latest review time on product r[i].p, Ip[x, y] is similar to that defined in Equation 4, but x and y are the posted time poinits of two reviews on the same product respectively, max(Ip) is the maximum time interval among all pairs of adjacent reviews on product r[i].p.

• **The repeatability measure** (F6)

Spammers might comment a product repeatedly. Therefore, we make a complement for F4 and F5 with checking whether r[i].u has cOlmnented the product r[i].p or not.

Sr = {l r[i].uEUp
o r[i].u rt. Up

where Up is the set of reviewers commented product r [i]. p.

## III. THE FAKE REVIEW DETECTION METHODS

A. **The supervised detection methods**

Based on the six features proposed above, we can transform the review r[i] into a vector Xi' Given a labelled review set {(Xl,yI),' " , (Xn,Yn)}, where Yi is the X;'s label. Based on these labeled samples, we can apply Logistic regression or SVM to predict the labels of reviews.

For the Logistic regression, the parameter vector w can be evaluated by Maximum Likelihood Estimations. Thus, the Logistic regression model for fake review detection can be defined as follows.

exp(w · X) P(y = LspamlX) = 1 + exp(w . X)
P(y 1 = LnormallX) = 1 + exp(w . X)

We also predict the review labels with the SVM, which tries to find a high-dimensional separating hyperplane between two groups of data. To simplify feature analysis in later section, we restrict our evaluation to linear SVM, which learns a weight vector wand bias term b, such that a review Xj can be classified by:

f) = sign(w . Xj + b)

It is worth noting that we need to update the corresponding centroids for Fl and F2, the signature sets for F3, the maximum time intervals for F4 and F5 and the comment status of each product for F6, after each review is processed.

B. **The threshold-based detection method**

Although the trained classifiers have stronger generalization ability. In practice, The fake reviews occupies a small proportion of all reviews. Thus, it would take some time to gather a certain amount of fake reviews for training. Until that, the fake reviews would make the negative influence.

Therefore, we devise a threshold-based solution for detecting the fake reviews without labeled samples. Recalled the features F 1-F6 discussed in Section II, each of them tries to highlight the fake reviews from different perspectives.

Intuitively, if a review is a spam, the sum of the feature values of FI-F6 would tend to be close to 6, since all of them locate in [0, 1] according to their definitions. Thus, we can evaluate the spam score of a review as following:

S(alSu + a2Sp + a3So + a4Su core j + a5Spj + a6Sr) = 6

Lk=l ak

where aI, a2,' .. , a6 are the weight parameters turning the contributions of feature F 1, F2, . . . , F6 separately. Since the spam score is normalized in [0,1]' we can determine whether a review is a spam with a predetermined threshold

T: { Lnormal SCOre r[i] < T, Lr[i] = Lspam SCOrer[i] > T

random SCOrer[i] = T

where the random means predicting r[i]'s label randomly.

## IV. EXPERIMENTS

### A. Dataset

One of the main challenges encountered by fake review detection is the absence of ground-truth, because fake reviews are difficult to find, even for human readers. Thus, we treat the duplicates/near-duplicates of reviews as the fake ones. This way is applied in some prior works like [1] and [4].

We reconstruct our fake review dataset with the following processes based on Liu's review dataset I , which was crawled from Amazon:

1) Removing the inactive products (review count < 10). 1 http://www.cs.uic.edu/ liubIFBS /sentiment-analysis.html

2) Sorting the reviews of each product with the displaying model "Newest First" on Amazon.

3) Calculating the Jaccard similarity for each pair reviews based on bigrams.

4) Treating the reviews as fake review candidates, whose Jaccard similarity is greater than or equal 0.7.

5) Sorting each pairs of candidates by their posted orders. We get some order copy chains. We remark the review at the head of each chain as the normal ones, because we can not be sure whether they are review spam.

We have collected 2000 fake reviews with above process together with the corresponding normal reviews based on the Liu's dataset. These reviews are sorted by their posted orders, which constitute an ordered review dataset. The reconstructed dataset includes 6824 products, 155080 reviews, 122672 reviewers and 2000 review spam.

### B. Experimental results and discussion

For the supervised solutions, the training samples should be chosen based on the reviews' posted orders. Thus, we can simulate the arrival orders of reviews in real-word, and avoid to use the later reviews to predict the front ones' labels.

In the first experiment, the count of training fake reviews is fix, and we compare the detection effects with different proportions of training fake reviews and normal ones by increasing the normal reviews' quantity. Namely, we use 50 fake reviews and 50 x i (2 :s; i :s; 20) normal ones for training, the remaining reviews are treated as the testing samples.

(FI (s» of Logistic regression and SVM for fake review detection. The detection effects of both methods are improved, when the normal review count is increased until it reaches 700. When we apply 50 fake reviews and 800 normal ones for training, the SVM achieves the highest FI(s) value (0.923), and the detection precision on spam ( pre c isions) is 0.939, the recall of spam (re calls) is 0.908. Moreover, we can observe that the SVM is not very sensitive to the normal review count.

For the contributions of the proposed features for fake review detection, Table I shows the effect of SVM with different feature options respectively, in which 50 fake reviews and 800 normal ones are used to train. According to the order from F 1 to F6, the "1" indicates the feature in corresponding location would be used. For example, the "11 0 Ill" means the feature F3 is excluded only.

the feature Fl, F2, F3, F5 and F6 are active for spotting fake review, since we remove any one of them will lead to the declining effect. Recalled the definition of F4 in Section II, it means that the smaller time interval between the posted time points of two adjacent reviews wrote by one reviewer, the greater probability of the review be fake. But the time granularity of review is the day in our dataset, while some normal reviewers also write multiple reviews in one day.

This may be the main reason that the feature F4 does not work. If we could obtain finer time granularity, we consider the feature F4 would make its contribution to fake review detection.

We can observe that the review content is more effective than the reviewer behavior for fake review detection. We consider the main reason still may be the time granularity.

The last experiment focuses on the threshold-based method. . Clearly, if a review is fake, its spam score would be close to l. We get the highest value of Fl (s) with the T = 0.55 and the feature weight setting "212111", which is very close to the best effect of the Logistic-based method. Moreover, with the increase of T, the pre c isions is increasing and the recalls is decreasing for the same set of parameters. Thus, we can tune the T to reach various effects for different application demands.

## V. CONCLUSION

In this paper, we explore the problem of identifying fake reviews from the review sequences, which consists of the ordered reviews according to their posted time points.

**THE EFFECTIVENESS OF THE THRESHOLD-BASED DETECTION METHOD**

Feature precisions recalls Fl(S) option accuracy Firstly, we highlight the fake reviews with six time sensitive features based on the review contents and the reviewer behaviors. Secondly, we devise the supervised solutions and a threshold-based one for spotting fake review respectively.

At last, we carry out a series of experiments on a real review set to verify the effects of the proposed methods, and analyze the utilities of the six proposed features in different cases. The experimental results show that the supervised methods can work with high precision and recall on fake reviews based on few training samples, and the threshold-based one can achieve a relatively good effect without training samples.

## REFERENCES

[1] N. Jindal and B. Liu. Analyzing and detecting review spam. In ICDM 2007, pages 547-552. IEEE, 2007.

[2] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. arXiv preprint arXiv:1107.4557, 2011.

[3] E.-P. Lim, y'-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw. Detecting product review spammers using rating behaviors. In CIKM 2010, pages 939-948. ACM, 2010.

[4] N. Jindal and B. Liu. Opinion spam and analysis. In WSDM 2008, pages 219-230. ACM, 2008.

[5] S. Xie, G. Wang, S. Lin, and P. S. Yu. Review spam detection via temporal pattern discovery. In SIGKDD 2012, pages 823-83l. ACM, 2012.

[6] G. Wang, S. Xie, B. Liu, and P. S. Yu. Identify online store review spammers via social review graph. ACM Transactions on Intelligent Systems and Technology (TIST), 3(4):61, 2012.

[7] A. Mukherjee, B. Liu, and N. Glance. Spotting fake reviewer groups in consumer reviews. In WWW 2012, pages 191-200. ACM, 2012.