ISSN: 2096-3246
Volume 52, Issue 02, May, 2020

https://doi.org/10.5281/zenodo.12706997

IDENTIFICATION OF FAKE REVIEWS WITH MACHINE LEARNING APPROACHES

VARADA DEEPTHI, ASONDI SREEPRADHA, AMRUTHA GOPA

Assistant Professor [1,2,3]

varadadeepthi@gmail.com, sreepradha5217@gmail.com, amruthacse9f@gmail.com
Department of Computer Science and Engineering, Sri Venkateswara Institute of Technology, N.H 44,
Hampapuram, Rapthadu, Anantapuramu, Andhra Pradesh 515722

Keywords:
Node Location, Cloud Safety, Data Deployment, and Throughput.

ABSTRACT

Nowadays, when choosing a brand, most of us look to online reviews. Unfortunately, review sites are increasingly facing the problem of opinion spam, which spreads false information with the intent of promoting or harming specific businesses through deceiving human readers or automated systems that analyse sentiment and opinion. It is for this reason that many data-driven methods for evaluating the veracity of user-generated material disseminated via social media in the shape of online reviews have been put forth in recent years. Reviewers, reviews, and the network structure that links various entities on the review-site in test are all aspects that different techniques take into account. The purpose of this article is to examine the most prominent review and reviewer-centric aspects that have been suggested in previous works as a means of identifying fraudulent reviews, with a focus on methods that make use of supervised machine learning. On the whole, these solutions outperform completely unsupervised methods, which often depend on graph-based techniques that take relational relationships in review sites into account, when it comes to overall performance. Also, several other novel elements that could be useful for distinguishing between real and false reviews are suggested and assessed in this study. With this goal in mind, we have developed a supervised classifier using Random Forests. Our model takes into account both established and novel characteristics, and it is trained on a massive labelled dataset. The usefulness of this investigation and the efficacy of the new characteristics in detecting singleton bogus reviews are shown by the positive findings that were achieved.

Introduction

A growing amount of user-generated content (UGC)—including textual, audiovisual, and video material—has been disseminated via the social Web and social media platforms. Thanks to Web 2.0 tools, anybody may publish anything on social media, almost without the need for a third party to verify its authenticity. This means that the credibility of the sources and the produced information cannot be pre-determined. Researchers are giving the problem of determining the veracity of information spread via social media platforms more and more attention in this context. Opinion spam, which spreads false information and has detrimental effects on both users and companies, has been the focus of extensive investigation on review sites. Here, the goal of opinion spam detection is to quickly recognise and label deceitful statements, phoney blogs, fake comments, fraudulent social network posts, and fake views [1]. Because user recommendations on review sites like TripAdvisor1 and Yelp 2 have such an impact on Websiteforadvice visitors, methods to detect false reviews have been developed specifically for these sites.That being said, recommending a service or product, like a restaurant,or a hotel when you rely on inaccurate information, it might lead to problems. So date, the majority of methods for identifying fraudulent reviews on these platforms have relied on supervised machine learning techniques and specific reviewer or review product properties. The use of these tools has been shown in research to effectively detect questionable material, reviewer actions, and, by extension, disinformation [2]. - ANewer methods propose supplementing traditional review sites with characteristics that take into account the underlying social structure of the network. These techniques often outperform supervised solutions, and they're commonly built on unsupervised graph-based methodologies. In contrast, supervised methods too have their share of problems. To begin, most existing methods have either taken a subset of attributes into account, or have tested on very tiny datasets retrieved from the famous review sites that were mentioned before. So, most of the time, the remedies that are suggested are either incomplete or depending on the review site. Taking into account the wide range of characteristics that have been suggested and used independently by supervised techniques, the objective of this article is to provide a feature analysis that demonstrates the best and most general features that can be utilised in the context of review sites to identify false reviews. Although some of these elements are familiar from previous works in the field, others are brand new and add to the paper's contribution.Using a supervised classifier based on a popular machine learning approach, we can assess how useful these variables are for distinguishing between real and false reviews. Regarding the literature, a general-purpose, publicly-available dataset from the Yelp.com review site has been taken into consideration. In terms of the impact of both individual features and sets of features, this enables us to provide more substantial outcomes. Specifically, it has become clear that a small set of characteristics significantly aids in determining the veracity of so-called singleton reviews. With these encouraging outcomes, it's clear that the feature analysis described in this article is both successful and potentially useful.

I. RELATED WORK

Researchers have offered a wide variety of methods to evaluate the veracity of information spread via social media in recent years, depending on the specifics of the situation [2]. Believability, trustworthiness, perceived dependability, competence, correctness, and a plethora of other ideas or mixes of these have long been linked to the notion of credibility [3]. Fogg and Tseng [4] state that there are several components that make to an information receiver's trustworthiness. The three main components of each piece of information are its origin, its structure and content, and the medium via which it is disseminated [5]. The effect of the communication medium on people's perceptions of information sources and content itself might alter their trustworthiness perceptions when taking these features into account [3, 5]. This is why it's crucial to address the topic of whether digital media bring additional elements that might affect credibility evaluation [6, 7]. When it comes to the social web, determining the reliability of information requires looking at user-generated material [8], author traits, and the features of social media platforms itself, which are the interconnections between the various entities. Depending on the context, these features can be basic linguistic aspects linked to the user-generated content (UGC) text, supplementary meta-data aspects linked to the review or tweet content, behavioural features derived from users' social media activity, or even linked to their profile (if available). Additionally, various methods have leveraged social features, which take advantage of the network structure and relationships connecting entities in social media platforms [9], [10], or product-based features, as in review sites devoted to products and services. 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] have been the most investigated tasks in the recent past pertaining to the automatic or semi-automatic assessment of the credibility of information on the Social Web. The bulk of these methods centre on data-driven approaches that use various models to categorise user-generated content (UGC) according to its trustworthiness.The most effective methods for detecting opinion spam, and more specifically for detecting fake reviews (the main focus of this paper), typically employ supervised or semi-supervised machine learning techniques that consider review- and reviewercentric features. The first methods relied only on language analysis, using unigrams and bigrams as primary textual characteristics derived from reviews [13, [14], [15], [16]. Various linguistic methods have put forth language model-based generative classifiers [17], [18]. Since it is extremely difficult for a human reader to differentiate between credible and untrustworthy information just by reading it, particularly in an era where the abilities to write false reviews are constantly improving [20], Mukherjee et al. [19] showed that relying solely on linguistic features is ineffective to identify fake reviews in real datasets.

Consequently, more effective multi-feature-based methods that use several characteristics of various types in addition to basic linguistic ones have been suggested. These methods include either by using MCDM (Multi-Criteria Decision Making) [22] or supervised or semi-supervised machine learning [1, 19, 21]. For assessment reasons, these methods often centre on tiny labelled datasets, which are often made up of data that is "near ground truth" [9]. The review site's network of entities (users, items, reviews, etc.) is often not taken into account, and qualities derived from these social links are often ignored. By contrast, graph-based techniques often make use of this kind of characteristic (together with the other traits mentioned earlier)[23], [24].Although supervised learning on a small set of categorization labels may supplement these later techniques, they are mostly unsupervised [25]. In comparison to supervised methods, completely unsupervised solutions often provide somewhat worse outcomes [2, 9, 20]. Taking supervised solutions into account, this paper reviews the literature on review- and reviewer-centric features that have been suggested for detecting fake reviews and discusses and analyses them on a broad level. It also suggests some new features that are well-suited to this purpose, especially for detecting singleton fake reviews, a problem that has not gotten the attention it needs. In order to circumvent the issue of the small size of the labelled datasets that have been addressed in previous work, two large-scale datasets that are publically accessible were used for assessment, as described in [25].

II. FEATUREANALYSIS

There are a lot of various factors that have been studied so far in the context of review sites to detect false reviews, as briefly mentioned in Section II. Feature classes have been treated independently by various methods in some instances. On the other hand, there are situations where the characteristics that were used are only a fraction of all the features that may be examined. To address unresolved difficulties, such how to identify singleton bogus reviews, new features can be suggested and evaluated. This is why we provide a high-level summary of all the characteristics that may be used to identify false views in this section.We take into account both the new characteristics suggested in this article and the important aspects gathered from the literature. Considering review- and reviewer-centric aspects, the two classes will be the most effective, as mentioned in the literature. can be seen in the sections that follow.Each section will explain the reasoning behind the selection of characteristics from the aforementioned classifications. The characteristics that are sourced from the literature will be referenced back to the original publication where they were first suggested. When a reference is missing, it means that the feature in question has been utilised by almost every conceivable approach. In conclusion, a feature that is first suggested in this article will be marked with the label [new] when it is present. A. Features Focused on Reviews Reviews make up the first category of attributes that have been taken into account. They may be retrieved from the review's text (textual features) or from the review's associated information (metadata features). Metadata about the reviewed company, including the date and time of publishing, and a rating (within a certain numerical interval), are standard features of all review sites. It is also important to pay close attention to metadata aspects related to the cardinality of a user's reviews. Actually, a significant portion of reviews are "singletons," meaning that reviewers only write one review each time period (i.e., there is only one review in the user account when the analysis is done). Designed features are necessary for this kind of evaluations. Actually, as we'll see later on, a lot of the attributes suggested in the literature are really based on data from several evaluations done by the same person. for dealing with one-offs, these characteristics become irrelevant for determining trustworthiness. It is, therefore, of the utmost importance to define appropriate characteristics that can effectively identify duplicate reviews. 1) Textual Features: Comparing phoney and real evaluations only based on their content is next to impossible, as briefly shown in Section II. The KL-divergence between spammers' and non-spammers' languages used on Yelp is quite minor, according to the research presented by Mukherjee et al. in [19]. Nevertheless, the positive outcomes achieved in [26] via the application of linguistic characteristics to a dataset that is specifically tailored to a certain domain (i.e., a Yelp dataset that only includes Japanese restaurants in New York) demonstrate that textual features may be beneficial, at least at a domain specific level. Statistical information and emotion assessments related to word use may be utilised as features in text extraction using Natural Language Processing methods. • Text: Multiple methods use unigrams and bigrams taken from review texts as textual characteristics, as shown in Section II. • Text statistics: Li et al. in [21] suggested several review content statistics as features: - The length of the opinion expressed in words; - Capitalization ratio, or the percentage of words in the review that are capitalised relative to the entire word count; - The percentage of words that are entirely capital letters, which includes terms with no lowercase characters; the frequency of first-person pronouns (such as "I," "mine," "my," etc.);

The percentage of sentences that finish in "exclamation" (i.e., with the sign "!"). • Perception assessments: - Subjectivity, or the ratio of subjective (expressing emotion or judgement) to objective (descriptive) language. 2) Meta-data features: these attributes may be derived from the meta-data associated with reviews or derived from the reviews' cardiovascularity in relation to the reviewer and the reviewed entity. • Essential details: - The entity's rating, expressed as a numerical number on a specified interval (e.g., 1–5 "stars"); - Rating deviation [27], which is the dispersion of the review's assessment from the entity's average rating; - Singleton [25], which denotes that the review is the only review published by a reviewer for a certain time frame (e.g., a day). Some straightforward and obvious heuristics are used by these fundamental aspects. In comparison to real reviews, fake reviews often have more "extreme" ratings, which means that the rating deviation from the entity's average rating is higher. What's more, a user who only posts one review could not be very active in the review site community, which could be a sign of unreliability.

A phenomenon known as "burst features" occurs when a large number of evaluations for an entity appear all at once. Both spam assaults and unexpectedly high popularity of the entities assessed may cause these review bursts. Review bursts tend to have similar characteristics, which makes it easy to spot clusters of fraudulent reviews by studying the burst's characteristics [28]. Two investigations on burst detection have been detailed in the literature [27], [28].This paper incorporates a number of aspects that take burstiness into account, drawing on the aforementioned works for inspiration.In relation to a certain reviewed entity, these qualities are associated with the time range in which the review was submitted. Review fraud is more likely to occur on days with an unusually high number of reviews and when there is a large disparity between the average rating of an entity in a review and its average rating (typically a drop from 3.5/5 to 2/5, for example) within a specific time frame. Although this assumption holds true for all review types, it has been much improved when it comes to detecting singleton phoney reviews, which are notoriously hard to spot without taking burstiness into account. In light of previous work, we have taken into account the following novel aspects:

- Density, which is the number of reviews for an entity on a specific day of publication; - Mean rating deviation, which is the deviation of an entity's average rating on that day from its average rating (in general); - Deviation from the local mean, which is the numerical evaluation of whether a given entity's rating in a review is close to the average rating on that day; - Early timeframe, which is the amount of time it takes before the first review on an entity is posted. Actually, spammers often review early in order to make their (false) comments more impactful on the audience.

B.Reviewer-centricFeatures

Features pertaining to the reviewers' actions make up this set of characteristics. This manner, we can take into account how consumers often write reviews rather than just the text and meta-data linked with them, which have their limitations when it comes to categorization. One approach is to use textual characteristics when thinking about reviewers' actions; this helps with the issue of duplicate reviews, which has been discussed and researched in a number of studies [1, 29, 30]. The following textual characteristics have been extracted from the source [31]: The following metrics are available: • Word number average, which measures the average number of words used by the user in her or his reviews; • Average content similarity, which measures the average similarity across all of the user's reviews; and • Maximum content similarity, which measures the evaluation of the maximum similarity across all of the user's reviews. A very high MCS will be achieved by users who copy and paste their reviews. Some may see this as an indication of questionable behaviour. One way to measure content similarity is by looking at the cosine similarity between the user's reviews represented by bag-of-words.
2) Rating Features: These features are built from an aggregate of data on the reviews and ratings that each reviewer has submitted: • The overall count of reviews The percentage of reviews that are unfavourable, positive, or "extreme" (those whose ratings fall on opposite ends of the scale) [27]; • Rating entropy[25], i.e., the entropyof rating distribution of user's (entity's) reviews; • Average deviation from entity's average [31], i.e., the assessment of whether a user's ratings given in her/his reviews are frequently extremely different from the mean of an entity's rating (much lower, for instance);

The squared departure of a user's rating from the ratings mean is known as the rating variance [new]. To better illustrate the distribution of ratings for a certain user, the variance rating tool has been implemented.
Thirdly, there are temporal features, which demonstrate the distribution of ratings over time and are

https://doi.org/10.5281/zenodo.12706997

based on actual data: • The user's activity time [31], which is the time difference between a reviewer's first and final review; • The date entropy [25], which is the average number of days that pass between successive pairs of reviews; • The maximum rating per day [25], which is the highest rating that a reviewer may submit on any given day; • Date variance [new], which is the squared variation of the timestamps of user reviews relative to the mean of timestamps.The distribution of reviews for a certain user across time frames may be better described with the addition of the variance as a time-related attribute.

III. SUPERVISEDCLASSIFICATION

To test how different features and sets of characteristics affect classification and overall performance, the classifier has been put into action. This is a crucial component of the study since it demonstrates the effects of various elements on a large-scale dataset, which makes it difficult to determine their relative value, as many publications only cover a selection of these traits and test them on little or review-site-dependent datasets.

IV. CONCLUSION

Research on methods for determining the veracity of content shared on social media, and more specifically, strategies for spotting opinion spam on review sites, has been growing in recent years. Typically, data-driven algorithms that take into account various reviewer, review, and social network characteristics that may be used to categorise reviews according to their reliability form the basis of approaches to fake review identification. In most cases, supervised classifiers are more effective and typically focus on aspects that are important to employees and reviewers. Most unsupervised classifiers employ graph-based models; they pay special attention to the social connections inherent in the review site for exams, among other factors. The main benefit of unsupervised solutions is that they do not need labelled datasets for training, but they are often less successful. On the other hand, supervised methods have shown their efficacy when applied to small or review-site dependent labelled datasets, as well as to generic subsets of characteristics. This article presents the results of a feature analysis that aims to summarise the key review- and reviewer-centric features that are suitable for detecting fake reviews. It also proposes new features that could be especially helpful for detecting singleton reviews, with a focus on the effectiveness of supervised classification. Our team has created a supervised classifier using Random Forests to assess the significance of these characteristics. A publicly accessible large-scale general-labeled dataset was used for assessment purposes to circumvent difficulties related to the scarcity of available ground facts. The usefulness of the suggested research is shown by the encouraging outcomes that were achieved.

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”,inAnIntegratedApproachtoCommunicationTheoryandResearch, 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 contemporarymediaenvironment,”AnnalsoftheInternationalCommunicationAssociation,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.

https://doi.org/10.5281/zenodo.12706997

[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 TechnologiesVolume 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 ngrams," 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 onE-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