IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 7, July 2021

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

|DOI:10.15680/IJIRSET.2021.1007031|

Sentiment-Enhanced Content-Based System for Online Recommendations and Rating Prediction

Sapana Supekar[1], Pratiksha Sarode[2], Manasi Rangate[3], Vaishnavi Sonavane[4],

Prof. Himanshu Joshi[5]

Department of Computer Engineering, Imperial College of Engineering and Research, Wagholi, Pune, India

ABSTRACT: generally the people trust on product on the basis of that product reviews and rating. People can remove a review allow to spammers to form spam studies about goods furthermore, administrations for different benefits. Recognizing these fake reviewers and the spam content is a big debated issue of research and despite of the way that a various number research has been done already. Up till now the ways set hardly differentiate spam reviews, and no one show the significance of every property type. In this investigation, a structure, named NetSpam, which uses spam features for demonstrating review data sets as heterogeneous information frameworks to design spam identification method into a group of issue in this network. Using the criticalness of spam features help we to obtain good outcomes regarding different metrics on review data sets. The commitment work is when client search question it will show all n-no of items just as suggestion of the item.

KEYWORDS: Fake Review, Machine Learning, social media, Social Network, Spammer, Spam Review

I. INTRODUCTION

In the dissemination of information online social media portals play an important role, and are often taken in the selection of goods and services as a key source of information for producers in their campaigns. In recent years, people depend in their dynamic practices on compiled audits, and they are promising/unpromising in their product and service range with positive/negative feedback. Written reports also allow service providers to demonstrate their offerings and services' standards. These reviews were therefore a significant factor when a market success was achieved while favorable reviews could lead to advantages for a company. The fact that someone of any personality will leave observations as an audit gives spammers a lovely opportunity to record falsified surveys to delude the conclusion of customers.

II. LITRETURE SURVEY

Ch. Xu et al.: The pair of wise features was first used to spot party colluders during spam campaigns for online product analysis and can expose complicity from a more fine-grained viewpoint in spam campaigns. A new Fraud Informer detecting framework is being proposed to deal with the intuitive and unmonitored, pair-wise functions. Benefits are: Pair smart functionality should be a rigorous paradigm for correlating colluders such that all website reviewers are positioned internationally so that top-class colder manipulated the perceived reputations of the objectives for their best interests. Benefit is a complex automating challenge.

G. Fei et al.: The paper proposes to create a network of reviewers in the form of a Markov Random Field (MRF) and apply the loopy believe propagation (LBP) approach to decide whether a reviewer is a spammer or not. A new evaluation tool for dynamically assessing observed spammers using their review classification. Benefits are: high precision, the approach suggested is efficient. To spot spammers in the examination of spammers. Automatically spot spammers. The downside is: A standardized spammer detection system is not used.

j. Minnich et al.: The problems in the paper are: to spot deceptive behavior, determine the reliability of revised websites, as some may have mis behavioral tactics, and develop successful revision aggregation solutions. The TrueView score in three separate versions proves that multi-site view synthesis provides the end user with important

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

|DOI:10.15680/IJIRSET.2021.1007031|

and functional details. Benefits include: Create new features that will easily distinguish cross-site inconsistencies, a hotel identification matching mechanism of 93\% precision. Enable the owner of the website to spot hotel mistakes. Activate confidential feedback for the end customer. Benefit is a complex automating challenge.

B. Viswanath et al.: In the paper unattended strategies for the identification of anomalies over user behavior are defined to differentiate likely bad conduct from usual conduct. To find fraudulent, corrupted and colluding identities with different intruder schemes without a prior mark while preserving low false positive rates. Detection of anomalies in Facebook advertisements to recognize anomaly. Reaches an identification rate of more than 66\% of misbehavior (over 94\ %) and less than 0.3\% false positives. The intruder tries to drain the advertiser's budget by clicking on advertisements.

Li, Z. et al.: It extending to a mutual positive and unlabeled learning algorithm called a multiple heterogeneous category classification (MHCC) (CPU). In the PU and non-PU learning environment, the proposed models will greatly improve the F1 scores of solid baselines. Benefits include: In PU and non-PU learning settings the proposed models will significantly improve the results of F1 from solid baselines. The model can be extended in other languages smoothly, using language-specific functions only. There are several likely incorrect feedbacks in the unlabeled collection that are found. Fake reviews conceal that Dianping's algorithm did not catch in unlabeled reviews. The ad-hoc consumer and IP labels that are used in MHCC cannot be very precise since they are calculated on adjacent summary labels.

M. Crawford et al.: This paper develops two different approaches to minimize the subset size of features in the spam region. The processes have filter-based rankers and word frequency selection functions. Benefits include: The first way is to choose the words most often appearing in the text easily. The second approach will use filter-based rankings to identify the characteristics and then pick the top features. Disadvantages are: Not all approaches that are often best fit a single scale.

H. Xue et al.: In the document it is possible for the people to regard feedback from people associated with them as more credible and to review spammers less likely to establish a broader relationship network with regular users to provide an accurate and productive way to identified revised spammers by adding social interaction assumptions. The benefits are: The suggested forecast based on confidence achieves greater precision than the conventional CF process. To solve the issue of sparsity and calculate the total confidence score for any device consumer used as a spam predictor. Benefits are: Analysis of required data collection.

E. D. Wahyuni et al.: In this paper, it is suggested that the text of an analysis can identify false feedback of a product. Briefly, the system suggested (ICF++) would calculate the integrity, the trustworthiness, and the durability of a commodity in an appraisal. Benefits include: Precision is stronger than ICF. The disadvantages are: process must be streamlined. Precision is optimizing.

R. Hassanzadeh et al.: This paper presents an outline of emerging problems in a variety of online social-network problem areas, which can be tackled through anomaly identification. It offers an outline of current anomaly detection methods and how those techniques are used for the study of the social network. Benefits include: Anomalies used to classify criminal acts are observed. Benefits are: The use of SNA anomaly detection methods needs to be enhanced.
R. Shebuti et al.: The paper offers a new comprehensive solution, called SpEagle, using information from both metadata (text, time stamp and rating), and relational data (network) to identify suspected consumers and ratings as well as spam-targeted items collectively under the single scheme. SpEagle employs a review-network-based classification task that embraces previous information, estimated by metadata, on the class distribution of nodes. The advantages are: when labelled data are usable, it allows smooth integration. It's very strong.

III. PROPOSED APPROACH

A new proposed framework consists in representing a set of reviews data provided as FIN (Fake Information Networks) and solving the issue of spam detection in a problem of FIN classification. In particular, to show the reviews data set as a FIN where the reviews are linked through different types of nodes (such as functionality and users). Then a weighting algorithm is used to calculate the importance (or weight) of each function. These weights are used to calculate the latest review labels using supervised and unsupervised procedures. Based on our observations, defining two views for features (review-user and behavioral-linguistic), the classified features as review behavioral have more

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

|DOI:10.15680/IJIRSET.2021.1007031|

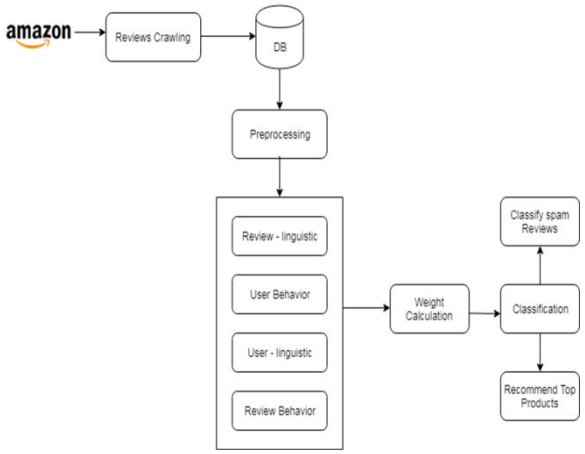
weights and yield better performance on spotting spam reviews in both semi-supervised and unsupervised approaches. The feature weights can be added or removed for labeling and hence time complexity can be scaled for a specific level of accuracy. Categorizing features in four major categories (review-behavioral, user-behavioral, review-linguistic, user-linguistic), helps us to understand how much each category of features is contributed to spam detection.

The general concept of our proposed framework is to model a given review dataset as a Heterogeneous Information Network and to map the problem of spam detection into a HIN classification problem. In particular, model review dataset as in which reviews are connected through different node types.

Advantages of Proposed System:

- It identifies spam and spammers as well as different type of analysis on this topic.
- Written reviews also help service providers to enhance the quality of their products and services.
- It identifies the spam user using positive and negative reviews in online social media.
- This framework displays only trusted reviews to the users.

Architecture:



Mathematical model for the low-level design (module-wise)

Spam Features:
User-Behavioral (UB) based features:

Burstiness: Spammers, usually write their spam reviews in short period of time for two reasons: first, because they want to impact readers and other users, and second because they are temporal users, they have to write as much as reviews they can in short time.

$$x_{BST}(i) = \begin{cases} 0 & (L_i - F_i) \notin (0,\tau) \\ 1 - \frac{L_t - F_t}{\tau} & (L_i - F_i) \in (0,\tau) \end{cases} \quad (1)$$

Where,
$L_i - F_i$ describes days between last and first review for $\tau = 28$.
Users with calculated value greater than 0.5 take value 1 and others take 0.

User-Linguistic (UL) based features:

Average Content Similarity, Maximum Content Similarity: Spammers, often write their reviews with same template and they prefer not to waste their time to write an original review. In result, they have similar reviews. Users have close calculated values take same values (in [0; 1]).

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

|DOI:10.15680/IJIRSET.2021.1007031|

Review-Behavioral (RB) based features:

- Early Time Frame: Spammers try to write their reviews a.s.a.p., in order to keep their review in the top reviews which other users visit them sooner.

$$x_{ETF}(i) = \begin{cases} 0 & (L_i - F_i) \notin (0, \delta) \\ 1 - \frac{L_t - F_t}{\delta} & (L_i - F_i) \in (0, \delta) \end{cases} \quad (2)$$

  Where,
  $L_i - F_i$ denotes days specified written review and first written review for a specific business. We have also $\delta = 7$. Users with calculated value greater than 0.5 takes value 1 and others take 0.
- Rate Deviation using threshold: Spammers, also tend to promote businesses they have contract with, so they rate these businesses with high scores. In result, there is high diversity in their given scores to different businesses which is the reason they have high variance and deviation.

$$x_{DEV}(i) = \begin{cases} 0 & \text{Otherwise} \\ 1 - \frac{r_{ij} - avg_{e \in E*j} r(e)}{4} > \beta_1 \end{cases} \quad (3)$$

  Where,
  β_1 is some threshold determined by recursive minimal entropy partitioning. Reviews are close to each other based on their calculated value, take same values (in [0; 1)).

Review-Linguistic (RL) based features:

Number of first-Person Pronouns, Ratio of Exclamation Sentences containing '!': First, studies show that spammers use second personal pronouns much more than first personal pronouns. In addition, spammers put'!' in their sentences as much as they can to increase impression on users and highlight their reviews among other ones. Reviews are close to each other based on their calculated value, take same values (in [0; 1]).

Algorithm Steps:

Input: review dataset, spam_feature_list, pre_labeled_reviews
Output: features importance (W), spamicity_probability (PR)
Step 1: u, v: review, y_u: spamicity probability of review u
Step 2: $f(x_{lu})$: initial probability of review u being spam
Step 3: P_l meta path based on feature l, L: features number
Step 4: n: number of reviews connected to a review
Step 5: m_u^{Pl} : the level of spam certainty
Step 5: $m_{u,v}^{Pl}$: the metapath value
Step 6: Prior Knowledge
Step 7: if semi-supervised mode
Step 8: if $u \in$ pre_labeled_reviews
Step 9: $y_u = label(u)$
Step 10: else
Step 11: $y_u = 0$
Step 12: else unsupervised mode
Step 13: $y_u = \frac{1}{L}\sum_{l=1}^{L} f(x_{lu})$
Step 14: Network Schema Definition
Step 15: schema = defining schema based on spam-feature-list
Step 16: Meta path Definition and Creation
Step 17: for $P_l \in$ schema
Step 18: for $u, v \in$ review_dataset
Step 19: $m_u^{Pl} = \frac{\lfloor s \times f(x_{lu}) \rfloor}{s}$
Step 20: $m_v^{Pl} = \frac{\lfloor s \times f(x_{lv}) \rfloor}{s}$
Step 21: if $m_u^{Pl} = m_v^{Pl}$
Step 22: $m_{u,v}^{Pl} = m_u^{Pl}$

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

|DOI:10.15680/IJIRSET.2021.1007031|

Step 23: else
Step 24: $m_{u,v}^{Pl} = 0$
Step 25: Classification - Weight Calculation
Step 26: for Pl ∈ schemes
Step 27: do $W_{Pl} = \frac{\sum_{r=1}^{n}\sum_{s=1}^{n} mp_{r,s}^{Pl} \times y_r \times y_s}{\sum_{r=1}^{n}\sum_{s=1}^{n} mp_{r,s}^{Pl}}$
Step 28: Classification - Labeling
Step 29: for u, v ∈ review_dataset
Step 30 $Pr_{u,v} = 1 - \prod_{Pl=1}^{L} 1 - m_{u,v}^{Pl} \times W_{Pl}$
Step 31: $Pr_u = avg(Pr_{u,1}, Pr_{u,2}, \ldots, Pr_{u,n}$
Step 32: return (W, PR)

IV. RESULT AND DISCUSSIONS

Experimental evaluation outcomes show the amazon API uses for product review dataset with higher percentage of spam reviews have better performance because when fraction of spam reviews increases, probability for a review to be a spam review increases and as a result more spam reviews will be labeled as spam reviews. The results of the dataset show all the four behavioral features are ranked as first features in the final overall weights. The Fig.2 graph shows the NetSpam framework features for the dataset have more weights and features for Review-based dataset stand in the second position. Third position belongs to User-based dataset and finally Item-based dataset has the minimum weights (for at least the four features with most weights).

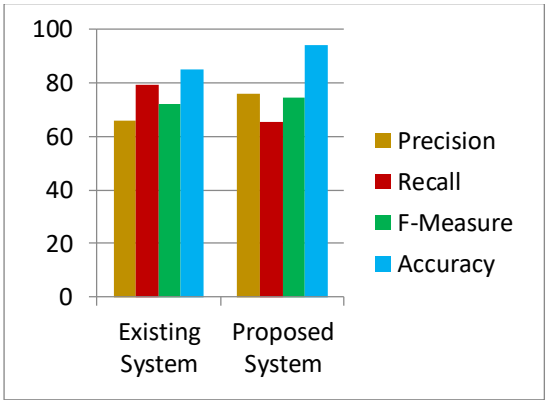


Fig. 2 Performance Analysis between existing and proposed system

V. CONCLUSION

In this proposed system investigation presents a novel in view of the met apathic concept and a graphical approach for naming feedback, this investigation proposes a novel spam identification scheme in particular SpamDup, based on the technique of ranking. The framework is validated with analysis datasets. The structure is introduced. Our insight shows that decided weights can be incredibly effective when detecting spam surveys and contribute to superior results when using this metaphorical notion. We have also found that SpamDup can sort out the importance of each feature even without a prepared range and that it can perform more efficiently throughout the process of expansion of highlights and superior to past works with only a few highlights. Furthermore, we show that, after the identification of four foundational classifications for highlights, the behavioral review ran higher than all other than AP, AUC and determined weight. The findings also affirm that the vast majority of the weighted highlights have no measurable effect, using different supervisory techniques, like the semi-managed approach, in the same way as in other datasets. This project is a contribution for the customer who receives the top-k product lists as well as one product recommendation object with a custom recommendation algorithm if searches are requested.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

|DOI:10.15680/IJIRSET.2021.1007031|

REFERENCES

[1]. Ch. Xu and J. Zhang," Combating product review spam campaigns via multiple heterogeneous pairwise features", In SIAM International Conference on Data Mining, 2014.
[2]. G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting bustiness in reviews for review spammer detection", In ICWSM, 2013.
[3]. A. j. Minnich, N. Chavoshi, A. Mueen, S. Luan, and M. Faloutsos, "True view: Harnessing the power of multiple review sites", In ACM WWW, 2015.
[4]. B. Viswanath, M. Ahmad Bashir, M. Crovella, S. Guah, K. P. Gummadi, B. Krishnamurthy, and A. Mis love, "Towards detecting anomalous user behavior in online social networks", In USENIX, 2014.
[5]. H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao," Spotting fake
reviews via collective PU learning", In ICDM, 2014.
[6]. M. Crawford, T. M. Khoshgoftaar, and J. D. Prusa," Reducing Feature Set Explosion to Facilitate Real-World Review Spam Detection", In Proceeding of 29th International Florida Artificial Intelligence Research Society Conference, 2016.
[7]. H. Xue, F. Li, H. Seo, and R. Pluretti," Trust-Aware Review Spam Detection", IEEE Trustcom/ISPA.,2015.
[8]. E. D. Wahyuni, A. Djunaidy," Fake Review Detection from a Product
Review Using Modified Method of Iterative Computation Framework", In Proceeding MATEC Web of Conferences, 2016.
[9]. R. Hassanzadeh," Anomaly Detection in Online Social Networks: Using Datamining Techniques and Fuzzy Logic", Queensland University of Technology, Nov, 2014.
[10] R. Shebuti, L. Akoglu," Collective opinion spam detection: bridging review networks and metadata", In ACM KDD, 2015.
[11] SaeedrezaShehnepoor, Mostafa Salehi*, Reza Farahbakhsh, Noel Crespi," NetSpam: a network-based spam detection framework for reviews in online social media", IEEE conference paper, 2017.
[12] G.D. Upadhyay, P. Barhate, "Classifying Handwritten Digit Recognition Using CNN and PSO", IJRTE, ISSN: 2277-3878, Volume-8 Issue-2, July 2019.
[13] G.D. Upadhyay, D. Pise, "Grading of Harvested Mangoes Quality and Maturity Based on Machine Learning Techniques", IEEE International conference on smart city and Emerging Technology,2018.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details