Opinion Spamming in Social Media: A Brief Systematic Review

Khairul Nizam Baharim[1,3] and Suraya Hamid[2]
[1]*University of Malaya, Malaysia, khairulnizam@siswa.um.edu.my*
[2]*University of Malaya, Malaysia, suraya_hamid@um.edu.my*
[3]*TM Research and Development, Malaysia, khairulnizam@tmrnd.com.my*

ABSTRACT

Opinion spamming in social media is an activity of people giving or sharing fake reviews or irrelevant opinions to online communities. The fake reviews are not merely misguided sentiment analysis and opinion mining system, but also severely affected online communities' decision and businesses reputation. Thus, opinion spamming detection (OSD) technique is needed to enhance an opinion mining system and prevent such cases from happening to the online communities. This study was conducted using the systematic literature review (SLR) procedure to classify known opinion spam features in social media platforms, and to reveal types of social media platforms that are being addressed by OSD's researchers. The result is, we found that, spatial and temporal factors in reviewer feature type is a current issue and is important to be solved because of spammer always changing their spamming strategy. On the other hand, most of the studies leveraged n-gram character and part-of-speech approaches in a review feature type because of its significant improved OSD's accuracy. Furthermore, we found that, most of the studies focused on trading and marketing-based social media platform, in which a lack of OSD's study in other forms of social media platforms i.e. social networking and user generated content sites.

Keywords: Opinion spamming detection, Opinion spam, Review spam, Fake reviews, Social media, Survey.

I INTRODUCTION

Social media are increasingly used by online communities and organizations in their daily decision-making. Online communities usually searched for an opinion of existing product consumers before purchasing new products or services. In the mean time, organization leveraged social media information to analyze and understand customer satisfaction and demand for future products development and services improvement. Because of that, sentiment analysis and opinion mining system now become more visible and freely accessible to the online community. For example, Google Shopping[2] and Bing Shopping[3] provide a review rating of the searched product, also a sentiment of product features related to it; where user could do a comparison across similar products before making the purchasing decision. Unfortunately, the sentiment analysis result may not accurate due to the possible existence of a fake review or an opinion spam.

In recent years, numerous high-profile fake review cases have been reported in the news media (Competition and Markets Authority, 2015; Griffith-Greene, 2014). Most of the cases involved businesses hiring people to write a fake review for them to promote their products and services. Unfortunately, it could be also to discredit their business competitors. Fake reviews in social media are thus not only harmful to consumers, but also to businesses. It would affect consumers' decision and businesses reputation severely.

Social media is a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0., and that allow the creation and exchange of user generated content (Kaplan & Haenlein, 2010). Definitely categorizing various types of social media platforms is impossible, but identifying their objectives is a key to understand how the platforms were built in different niches. Dijck (2013) defines general types of social media platform as follows:

- Social networking sites (SNS) – These sites primarily promote interpersonal contact, whether between individuals or groups of people. It allows personal, professional and geographical connections exchange. Examples are Facebook, Twitter, LinkedIn, Google+, and Foursquare.
- User generated content sites (UGC) – These sites support creativity, foreground cultural activity, and promote the exchange of amateur or professional content. Well-known UGC sites are YouTube, Blogger, WordPress, and Wikipedia.
- Trading and marketing sites (TMS) – These sites principally aim at exchanging products or selling them. TMS usually contain product reviews by the consumers. Amazon and eBay come to mind as notable examples.

[2] https://www.google.com/shopping

[3] http://www.bing.com

- Play and games sites (PGS) – These sites provide online gamers for interaction. Popular games such as a FarmVille, CityVille, The Sims Social, allow online users to communicate and exchange games feature.

However, there were no exact boundaries separating the social media platforms (Dijck, 2013). For example, SNS and TMS sites could also have creative content generated by users i.e. UGC. Thus, in this study, we scoped our review of OSD in social media for particular SNS, UGC and TMS platforms.

The contribution of this study is as follows:

- It has discovered and confirmed type of social media that are being addressed by OSD's researchers; where most of OSD's study were focused on TMS category and lack of OSD's study in other forms of social media platform category i.e. SNS, UGC.
- It has revealed and tabulates type of OSD features in social media platform reported by the researchers. This findings complement with the latest OSD's survey in (Heydari et al., 2015). The latest opinion spam features being studied were related to reviewer behavior feature and spatial-temporal factors.

The remainder of this work is structured as follows: In section II, we describe our review methodology and present the result and discussion in section III. In section IV, we conclude this study and propose an avenue of future work.

II METHOD

The SLR procedure (Kitchenham, 2004; Kitchenham et al., 2009) was first published in software engineering domain. Lately, it has been used widely in various software related domains such as information systems, computer networks, and mobile application. Hence, this study used SLR procedure to review the state-of-the-art in opinion spam detection research, particularly in social media platforms.

A. Research Questions

The research questions that addressed by this study were:

RQ1. What type of social media platforms were being addressed by OSD's researchers?
RQ2. What types of opinion spam features in social media were being used by OSD's researchers?

With respect to RQ1, opinion spam problem was first formulated by Jindal & Liu (2007) in the context of product reviews in Amazon platform, which is a type of TMS. Further comprehensive opinion spam analysis continued in (Jindal & Liu, 2008). Since then, OSD are mostly studied in the context of online reviews and not much study has been done in the contexts of other forms (e.g. forum discussions, blogs, microblogs) of social media (Liu, 2015). To address RQ1, we identified OSD's study published each year, the quality of journal/conferences that published them and scope of the study or dataset that are being used.

With respect to RQ2, Heydari et al. (2015) highlighted the issue of extracting the most effective and efficient OSD's features reported in literatures. To address RQ2, we identified empirical OSD's literature in social media, then captured and classified the reported OSD's features.

B. Research Questions

The search process was a manual search using two most important free citation-based academic search engines i.e. Google Scholar[4] and Microsoft Academic[5]. The search date range was set between 2007 and 2016, as the leading article by Jindal & Liu (2007) was published after Oct 31st, 2007.

The search keywords grown during the search process as depicted in Figure 1. It started by using well-known relevant keywords (we called it "seed keywords") extracted from (Jindal & Liu, 2008) article as follow: "opinion spam", "review spam", "fake reviews". A manual search was performed then using OR/AND Boolean operations. For example, the search command is: "opinion spam" OR "review spam" OR "fake reviews". The article was selected by the researcher based on its relevant title, keywords, and abstract. New relevant keywords found in the selected article were used in the next round of search until no new result appeared.

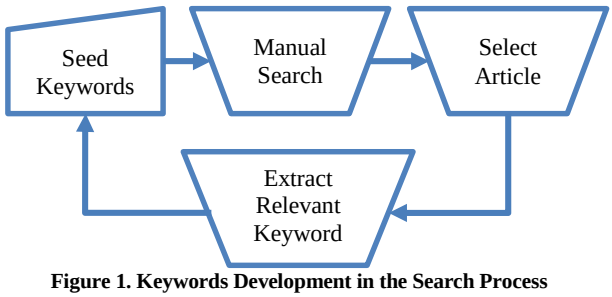

Figure 1. Keywords Development in the Search Process

The selected articles from the search process were then filtered by its quality. Relevant data were collected and analyzed to answer the research questions. The following sections detailed the process after the articles were selected.

[4] https://scholar.google.com/
[5] https://academic.microsoft.com/

C. Quality assessment

Articles on the following topics were excluded:

- Non-related article based on its title, keywords or abstract (not related with a problem of opinion spam in social media).
- Duplicate articles of the same study (when several articles of a study exist in different journals, the most complete version of the study was included in the review).
- Non-empirical studies (because we wanted to extract used OSD's features).
- Informal empirical studies (no defined methodology, dataset, and finding result).

D. Data collection

The data extracted from each selected article were:

- Authors.
- Article's year and keywords.
- The source (journal or conference).
- Other indexed source e.g. Web Of Science (WoS).
- Form of social media or dataset.
- OSD's features.

E. Data analysis

The data was tabulated to show:

- Literatures quality.
- Type of social media platforms (addressing RQ1).
- Classification of OSD's features (addressing RQ2).

III RESULTS & DISCUSSION

Due to space limitations, we only tabulate most significant articles based on their citation and organization reputation. The list of searched articles depicted in Table 1 for literature quality assessment. The "Selected Article" column in Table 1 indicates the articles that we have used to produce results in Table 2 for addressing RQ1, Table 3, Table 4, and Table 5 for addressing RQ2.

Table 1. Literatures Quality Assessment

Author(s)	Date	Source	Main Indexed Sources	Duplicate with Article	Method-ology	Data set	Empirical Result	Selected Article
(Jindal & Liu, 2007)	2007	Conf. ICDM	WoS, IEEE	-	Y	Y	Y	N
(Jindal & Liu, 2008)	2008	Conf. WSDM	ACM	(Jindal & Liu, 2007)	Y	Y	Y	Y
(Lim, Nguyen, Jindal, Liu, & Lauw, 2010)	2010	Conf. CIKM	ACM	-	Y	Y	Y	Y
(Jindal, Morgan, & Liu, 2010)	2010	Conf. CIKM	ACM	-	Y	Y	Y	Y
(F. Li, Huang, Yang, & Zhu, 2011)	2011	Conf. IJCAI	ACM	-	Y	Y	Y	Y
(Ott, Choi, Cardie, & Hancock, 2011)	2011	Meeting ACL	ACM	-	Y	Y	Y	Y
(Wang, Xie, Liu, & Yu, 2011)	2011	Conf. ICDM	IEEE	-	Y	Y	Y	Y
(Arjun Mukherjee, Liu, & Glance, 2012)	2012	Conf. WWW	ACM	-	Y	Y	Y	Y
(Fei et al., 2013)	2013	Conf. ICWSM	ACM	-	Y	Y	Y	Y
(H. Li, Liu, Mukherjee, & Shao, 2014)	2014	Journal	WoS	-	Y	Y	Y	Y
(Banerjee & Chua, 2014)	2014	Conf. SAI	WoS IEEE	-	Y	Y	Y	Y
(H. Li, Chen, Mukherjee, Liu, & Shao, 2015)	2015	Conf. ICWSM	AAAI	-	Y	Y	Y	Y
(KC & Murkherjee, 2016)	2016	Conf. WWW	ACM	-	Y	Y	Y	Y

Our findings in Table 2 shows that most of the OSD's studies were related to online review sites, particularly in TMS-based social media platforms. It confirmed the highlighted issue in RQ1. The social media platforms that are being addressed were: Amazon, Epinions, Dianping, TripAdvisor, and ResellerRatings.

The main obstacle in OSD study was to find or build gold-standard opinion spam dataset in order to evaluate OSD's technique in those platforms. H. Li et

al. (2015) seems had a large-scale labeled fake reviews dataset, but the data was private due to confidential agreement with Dianping. The only available small-size public dataset[6] for OSD modeling was created by (Ott, Cardie, & Hancock, 2012), particularly for TripAdvisor platform.

Table 2. OSD by Social Media Category and Platform

Social Media Category	Platform/ Dataset	Author(s)
Trading & marketing sites (TMS)	Amazon	(Fei et al., 2013; Jindal & Liu, 2008; Jindal et al., 2010; Lim et al., 2010; Arjun Mukherjee et al., 2012)
	Epinions	(F. Li et al., 2011)
	TripAdvisor	(Banerjee & Chua, 2014; Ott et al., 2011)
	ResellerRatings	(Wang et al., 2011)
	Dianping	(H. Li, Chen, et al., 2015; H. Li et al., 2014)
	Yelp	(KC & Murkherjee, 2016)

In the context of product reviews, there were three main types of reviews (Jindal & Liu, 2007, 2008):

- Type 1 (untruthful opinion) – It is a false opinion to lead the readers to positive or negative sentiment of the product.
- Type 2 (reviews on brand only) – Such reviews did not comment the product itself; instead emphasize the seller, organization or business.
- Type 3 (non-reviews) – Such reviews did not contain opinions, thus did not serve the purpose of reviews. It can be categorized into two main sub-categories: (1) Advertisements and (2) Other type of non-reviews such as question-and-answer communication between seller and reviewer.

Jindal & Liu (2007, 2008) considers duplicate and near-duplicate reviews as Type 1 reviews, which is one of opinion spamming factors that could be used for building OSD's model. Later in OSD's study, there were various complex opinion spamming scenarios and features identified by OSD's researchers.

We used general type of OSD's features category that were defined in (Jindal & Liu, 2007, 2008) to classify the collected OSD's features. Those were: (1) Review Features, (2) Reviewer Features, and (3) Product Features. However, later studies focused on reviewer behavior, in which we classified it as a kind of Reviewer Features in Table 4. Arjun Mukherjee et al. (2012) categorized spamming reviewer behavior indicators as: (1) Group Spam Behavior, and (2) Individual Spam Behavior. Rich reviewer behavior indicator further experimented in (H. Li, Chen, et al., 2015) were related to spatial and temporal features. The experimental result shown, by combining all kind of features that were behavior (A Mukherjee, Venkataraman, Liu, & Glance, 2013), linguistic (Ott et al., 2011) and spatial-temporal increased the accuracy of OSD's technique (H. Li, Chen, et al., 2015).

We classified type of OSD's features depicted in Table 3, Table 4, and Table 5 to answer the RQ2. In review features, n-gram characters and part-of-speech approaches were mostly used because of its significant improved OSD's accuracy. As discussed earlier, spatial and temporal factors in reviewer features are the current issues that are being explored by OSD's researchers. It is important because of professional opinion spammers always change their strategy in order to gain business profit.

Table 3. Review Features-based OSD

Feature(s)	Author(s)
Metadata – e.g. total-feedback, helpful feedback, title-length, body-length, review-position	(Jindal & Liu, 2008; F. Li et al., 2011)
Textual – e.g. capital, numeral, personal-pronouns, question, exclamation	(Jindal & Liu, 2008; F. Li et al., 2011; Arjun Mukherjee et al., 2012)
Similarity - e.g. similar-with-other-reviews	(F. Li et al., 2011)
Rating – e.g. review-rating, deviation-average, feature-rating, after-good/bad review?	(Jindal & Liu, 2008; Jindal et al., 2010; F. Li et al., 2011)
Sentiment analysis	(Jindal & Liu, 2008; KC & Murkherjee, 2016; F. Li et al., 2011)
N-gram characters	(Jindal & Liu, 2008; KC & Murkherjee, 2016; F. Li et al., 2011; H. Li, Chen, et al., 2015; H. Li et al., 2014; Ott et al., 2011)
Part-of-speech	(Banerjee & Chua, 2014; H. Li, Mukherjee, Liu, Kornfield, & Emery, 2015; Arjun Mukherjee et al., 2012; Ott et al., 2011)
Psycholinguistic – using Linguistic Inquiry and Word Count (LIWC).	(Banerjee & Chua, 2014; H. Li, Mukherjee, et al., 2015; Ott et al., 2011)
Readability – e.g. complexity, reading-difficulty	(Banerjee & Chua, 2014)
Honesty – store-reliability, agreement-with other-reviewer-within-time-window.	(Wang et al., 2011)

[6] http://myleott.com/op_spam/

Table 4. Reviewer Features-based OSD

Feature(s)	Author(s)
Review – e.g. wrote-first-review, the-only reviewer, multi-review-single-product, multi-review-group-product, review-diff-brand, burst-review-ratio, similar-review-diff-product, review-on-weekend, posted-via-PC,	(Fei et al., 2013; Jindal & Liu, 2008; F. Li et al., 2011; H. Li, Mukherjee, et al., 2015; Lim et al., 2010; Arjun Mukherjee et al., 2012)
Rating – e.g. avg/stdev-rating-given, good/bad-rating-given, deviation-avg-rating, weight-early-rating, diff-brand-diff rating,	(Fei et al., 2013; Jindal & Liu, 2008; F. Li et al., 2011; H. Li, Mukherjee, et al., 2015; Lim et al., 2010; Arjun Mukherjee et al., 2012)
Profile – e.g. reviewer-id, real-name?, homepage?, self-description, rank-popularity, registered-user,	(Jindal et al., 2010; F. Li et al., 2011; H. Li, Mukherjee, et al., 2015)
Trustworthy – e.g. reviewer-trust-reviewer, high-honest-review-score, amazon-verified-purchase,	(Fei et al., 2013; F. Li et al., 2011; Wang et al., 2011)
Group behavior – e.g. group-time-window, group-deviation, group-content-similar, group-early-time, group-size, group total-product,	(Arjun Mukherjee et al., 2012)
Location – e.g. user-distance, avg-travel-speed, avg-distance, unique-IP, unique-cookies, unique-cities-writing-review,	(H. Li, Mukherjee, et al., 2015)

Table 5. Product Features-based OSD

Feature(s)	Author(s)
Price	(Jindal & Liu, 2008)
Sales – e.g. sales-rank	(Jindal & Liu, 2008)
Rating – e.g. product-rating, avg/stdev product-rating	(Jindal & Liu, 2008; F. Li et al., 2011)
Profile – e.g. product-id, brand-id	(Jindal et al., 2010)
Review – e.g. brand/product-mentioned, review similar with product features, first product-review?	(F. Li et al., 2011)
Reliability – e.g. trustworthy-reviewer-say good	(Wang et al., 2011)

IV CONCLUSION

Opinion spamming in social media is a critical problem that needs to be solved because of its impact towards consumers and businesses decision. In this study, we tabulated a list of significant OSD articles from 2007 till early 2016 to show the known opinion spam features in OSD and social media that are being addressed. The findings confirmed that most of the OSD studies are in online reviews platform or TMS social media category. The latest opinion spam feature being studied is related to reviewer behavior and spatial-temporal features. Our future works are to explore OSD in social networking sites and user generated content platforms. We will perform empirical studies to discover the most effective and efficient OSD features.

REFERENCES

Banerjee, S., & Chua, A. Y. K. (2014). Applauses in hotel reviews: Genuine or deceptive? *Proceedings of 2014 Science and Information Conference, SAI 2014*, 938–942. http://doi.org/10.1109/SAI.2014.6918299

Competition and Markets Authority. (2015). *CMA acts to maintain trust in online reviews and endorsements*. Retrieved from https://www.gov.uk/government/news/cma-acts-to-maintain-trust-in-online-reviews-and-endorsements

Dijck, J. Van. (2013). *The culture of connectivity: A critical history of social media*. New York: Oxford University Press.

Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2013). Exploiting Burstiness in Reviews for Review Spammer Detection. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 175–184.

Griffith-Greene, M. (2014, November 6). Fake online reviews: 4 ways companies can deceive you. *CBC News*. Retrieved from http://www.cbc.ca/news/business/fake-online-reviews-4-ways-companies-can-deceive-you-1.2825080

Heydari, A., Tavakoli, M. A., Salim, N., & Heydari, Z. (2015). Detection of review spam: A survey. *Expert Systems with Applications*, *42*(7), 3634–3642. http://doi.org/10.1016/j.eswa.2014.12.029

Jindal, N., & Liu, B. (2007). Analyzing and detecting review spam. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 547–552. http://doi.org/10.1109/ICDM.2007.68

Jindal, N., & Liu, B. (2008). Opinion spam and analysis. *Proceedings of the International Conference on Web Search and Web Data Mining WSDM 08*, 219. http://doi.org/10.1145/1341531.1341560

Jindal, N., Morgan, S., & Liu, B. (2010). Finding Unusual Review Patterns Using Unexpected Rules. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 1549–1552). http://doi.org/10.1145/1871437.1871669

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68.

KC, S., & Murkherjee, A. (2016). On the Temporal Dynamics of Opinion Spamming : Case Studies on Yelp. In *International World Wide Web Conference Committee (IW3C2)* (pp. 369–379). http://doi.org/10.1145/2872427.2883087

Kitchenham, B. (2004). *Procedures for Performing Systematic Reviews*. *Joint Technical Report*.

Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering - A systematic literature review. *Information and Software Technology*, *51*(1), 7–15. http://doi.org/10.1016/j.infsof.2008.09.009

Li, F., Huang, M., Yang, Y., & Zhu, X. (2011). Learning to identify review spam. In *International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 2488–2493). http://doi.org/10.5591/978-1-57735-516-8/IJCAI11-414

Li, H., Chen, Z., Mukherjee, A., Liu, B., & Shao, J. (2015). Analyzing and Detecting Opinion Spam on a Large-scale Dataset via Temporal and Spatial Patterns. *Proceedings of Ninth International AAAI Conference on Web and Social Media*, (MAY), 634–637.

Li, H., Liu, B., Mukherjee, A., & Shao, J. (2014). Spotting Fake Reviews using Positive-Unlabeled Learning. *Computación Y Sistemas*, *18*(3), 2–10. http://doi.org/https://dx.doi.org/10.13053/CyS-18-3-2035

Li, H., Mukherjee, A., Liu, B., Kornfield, R., & Emery, S. (2015). Detecting Campaign Promoters on Twitter Using Markov Random Fields. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 290–299. http://doi.org/10.1109/ICDM.2014.59

Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., & Lauw, H. W. (2010).

Detecting Product Review Spammers using Rating Behaviors. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 939–948. http://doi.org/10.1145/1871437.1871557

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments and Emotions*. New York: Cambridge University Press.

Mukherjee, A., Liu, B., & Glance, N. (2012). Spotting fake reviewer groups in consumer reviews. *Proceeding WWW '12 Proceedings of the 21st International Conference on World Wide Web*, 191–200. http://doi.org/10.1145/2187836.2187863

Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013). What yelp fake review filter might be doing? In *ICWSM*.

Ott, M., Cardie, C., & Hancock, J. (2012). Estimating the prevalence of deception in online review communities. *Proceedings of the 21st International Conference on World Wide Web - WWW '12*, 201–210. http://doi.org/10.1145/2187836.2187864

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding Deceptive Opinion Spam by Any Stretch of the Imagination. *Proceedings of the 49th Annua Meeting of the Association for Computational Linguistics*, 11.

Wang, G., Xie, S., Liu, B., & Yu, P. S. (2011). Review graph based online store review spammer detection. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 1242–1247. http://doi.org/10.1109/ICDM.2011.124