International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 3 (2018) pp. 1628-1640

Group-Author Model for Latent Social Astroturfers Group Detection

Noora Alallaq [a] , Muhmmad Al-khiza'ay [b], Ahmed Al-Sudani[c] , N. Maheswari[d] and Ahmed Almansoori[e]

[abce] *School of Information Technology, Deakin University, Burwood, 3125, Australia.*
Email: { nalallaq, malkhiza, aralsuda, ajalma}@deakin.edu.au
[d] *school of computing science and engineering ,VIT university ,Chennai 600127, India*
Email: maheswari.n@vit.ac.in

Abstract

Online reviews play vital role in influencing people while making decisions in various sectors. However, in the wake of recent news, it is understood that astroturfing reviews or fake reviews caused unpleasant manipulations in the decisions of people. In this backdrop, astroturfing detection has attracted both industry and academia. Even in the presence of prevailing astroturfing, most of the review sites do not provide effective filters for astroturfing reviews. Unfiltered reviews can be deceptive and may promote or discredit certain product or service. Group of individuals who are intentionally spreading opinion either positively or negatively on a chosen service or product in specific period and in an organized fashion is known as astroturfing group. The existing literature focused on topic models and author models using Latent Dirichlet Allocation (LDA). However, there is no model which reflects both author groups that stay latent in the opinion manipulation with astroturfing reviews. In this paper we proposed and implemented a Group-Author model which is based on modified LDA with underlying unsupervised learning method known as clustering technique for text mining. We proposed a variant of LDA which is known as Group-Author model which takes two global parameters pertaining to corpus. They are known as author distributions in corpus α and astroturfing group author distributions in corpus β . We defined two algorithms namely Latent Astroturfer Group Detection (LAGD) and Temporal Filtering Algorithm (TFA) for discovering social astroturfing groups which are latent and validating such groups respectively. We built a prototype application to demonstrate proof of the concept. The empirical results revealed the utility of the proposed model in terms of discovering astroturfing groups besides reducing time and space complexity.

Keywords: Reviews, astroturfing, astroturfing groups, astroturfing group detection, unsupervised learning

INTRODUCTION

Astroturfing is the process of making fake online reviews where a group of individuals participate in a scheduled campaign in order to promote or demote a product or service. Due to emergence of social media, people of all walks of life are easily connected with virtual communities over Internet. This is the reason why people are being influenced by online reviews (social mirror effect). User generated online reviews started showing significant impact on people and play vital role in success or failure of a product or service. There is no much problem with this. However, the problem is with opinion spammers who provide fraudulent reviews [1]. Social web sites that facilitate online reviews such as YELP.com have certain mechanisms to prevent fraudulent reviews or astruturfing reviews. Group of individuals behind the intentional and fraudulent campaign of astroturfing is known as astroturfing group which is latent unless it is mined or discovered effectively. Though YELP.com is providing its proprietary method for preventing astroturfing, the research revealed that it is still inadequate to control opinion spammers [2].

Many researchers contributed towards dealing with opinion spam pertaining to online reviews. Out of them some important ones are described here. FraudEagle proposed by Akoglu et al. [16] detect fake reviews and provide scores for them. Metzger *et al.* [20] explored credibility evaluation online using social and heuristic approaches. According to Luka and Zervas [21] it is found that 16% of reviews made in YELP are filtered out as they are more extreme in nature. Restaurants those are weak in reputation committed review fraud. Lolaus is an online content rating system proposed by Kahoka et al. [22] to control malicious reviewers. NetSpam is yet another tool proposed by Shehnepoor et al. [24] for detecting spam reviews over social media.

From the literature it is found that more research is required to have efficient control of astroturfing. Many solutions found were on the Latent Dirichlent Allocation (LDA) model in the form of either author model or topic model for effective text mining. With respect to astroturfing two important distributions are found in the astroturfing reviews. They are known as distribution of authors and distribution of astroturfing author groups that are latent. As Group-Author model is not found in the literature, for the first to time to our knowledge, discovering latent social astroturfing groups. This is the motivation behind the research in this paper which addresses the problem by using Group-Author model which can adequately model the two latent distributions in the corpus of online reviews. Our contributions in this paper are as follows.

- We proposed Group-Author model for latent astroturfing group detection from online review corpus. This model is based on a modified LDA model which uses two parameters namely author distributions (α) and latent astroturfing group distribution (β).
- We proposed two algorithms namely Latent Astroturfer Group Detection (LAGT) and Temporal Filtering Algorithm (TFA). The former is used to extract latent astroturfing groups while the later is

used to validate the groups and updating them with time window behaviour of astroturfing groups.

- We collected 50 datasets with online reviews from YELP.COM. Each dataset reveals reviews pertaining to a restaurant.
- We built a prototype application to demonstrate proof of the concept. Our experimental results reveal the utility of the proposed methodology.

The remainder of the paper is structured as follows. Section 2 reviews literature on various approaches for filtering or finding fake reviews given by online users. Section 3 provides important definitions and formulates the problem addressed. Section 4 presents the proposed methodology covering author model, group-author model and various details about astroturfing group detection by modelling the behaviour of astroturfing groups that produce astroturfing reviews with certain procedure. Section 5 describes the experimental design used to evaluate the proposed methodology. Section 6 presents experimental results including astroturfing group detection using group-author model, time and space complexity. Section 7 evaluates the results with performance metrics such as precision and recall. Section 8 discusses threats to validity of the proposed methodology. Section 9 draws conclusions besides providing directions for future work.

RELATED WORKS

Akoglu *et al.* [16] proposed a framework known as FraudEagle to detect fake reviews in online review datasets. It analyzes network effects pertaining to products and reviewers. It provides scoring to reviews and users for effective fraud detection. It works with unsupervised learning method and scalable for large datasets. In the wake of security threats over social networks, Fire *et al.* [17] made a review of security and privacy risks in social networks. They focused on the threats that come due to fake users and their behaviour with other users especially children. They found threats to children such as cyber bullying, risky behaviours, and online predators. Mukherjee *et al.* [18] studied the filtering mechanism provided by YELP which is one of the web sites where online reviews are made. They found that YELP has an effective filtering mechanism that can take care of abnormal spamming behaviours.

Malbon [19] explored the consequences of taking fake online reviews seriously. Besides it argues that policy makers need to take fake reviews seriously. As consumer reviews can affect marketplace, it is essential to ensure that genuine reviews prevail in the review web sites and help consumers to make well informed decisions. The rationale behind this is to have control over information asymmetry over online social networks (OSNs). Metzger *et al.* [20] explored credibility evaluation online using social and heuristic approaches. They found that most of the people make credibility assessment on other people through some other people only. Credibility makes the data to be used in heuristic processes in order to have good decision making.

Luka and Zervas [21] from Harvard business school investigated consumer reviews available with YELP. They studied review fraud and economic incentives pertaining to the fraud. They studied the reviews that are suspected by YELP's filtering algorithm. They found that 16% of reviews made in YELP are filtered out as they are more extreme in nature. Restarants those are weak in reputation committed review fraud. There are chain restaurants that do not get benefited much from YELP generally do not commit review fraud. Kahoka *et al.* [22] proposed an online content rating system known as lolaus which is used to overcome malicious review activities being practiced by fraudsters using web sites like YELP. They used two techniques to overcome malicious behaviours. Weighing ratings and relative ratings in order to defect the system from multiple identity attacks and reduce the effect of bought ratings respectively. Cachia *et al.* [23] explored the potential of online social networks and their effect on the society. They found OSNs very useful for creativity, extracting emerging social behaviour, and garnering collaborative intelligence.

Shehnepoor *et al.* [24] proposed a framework named NetSpam to detect spam reviews over social media. The tool makes use of spam datasets and model a system which can detect spam messages. The model gets updated with new patterns of spam. NetSpam works well in four features such as user-linguistic, review linguistic, user-behavioural, and review-behavioural. Mukherjee *et al.* [25] proposed a mechanism for identifying pseudo reviews. They used Amazon Mechanical Turk (AMT) to generate fake reviews. They found that YELP's filter is only 67.8% accurate and such filtering systems are to be improved further. Anderson and Magruder [26] studied effects of online review databases in the real world. They proposed a regression discontinuity design for estimating the consequences of ratings. With reviews they found restaurants could increase their profit around 49% more. Zhang *et al.* [27] studied privacy and security of OSNs. They identified design conflicts and presented many useful insights in terms of opportunities and challenges.

Lampe *et al.* [28] proposed a theoretical framework related to profiles in social networking sites and focused particularly on transaction cost theory, common ground theory, and signalling theory. These are used for predicting profile elements in OSNs. Sirivianos *et al.* [29] proposed a framework known as FaceTrust for assessing credibility of people online over OSN. They found utility of anonymity over OSN besides security threats of the same. Online scams occur to exploit naive users of Internet which can be minimized with FaceTrust. Bello-Orgaz *et al.* [30] studied social big data and issues and challenges in processing such data. They opined that social big data throws security and privacy challenges as well. Pallis *et al.* [31] explored the present status of OSN and its future trends. Zhou *et al.* [32] on the other hand investigated the role of OSN in personalized recommendations. Again here they cautioned about fake users and fake information.

Lai *et al.* [33] proposed a novel review span detection method that is able to identify spam reviews. They developed an association mining method known as high-order concept association mining for detecting untruthful reviews. It was based on inferential language model. The model does not depend explicitly on review features. Moretti *et al.* [34]

opined that fake news and social media manipulated or affected the results of 2016 elections in USA. Fake news stories that are widely spread over social networks caused issues as the people believe them as real and make their decisions. Goga [35] studied social networks and found many insights such as privacy issues, security threats, and fake user accounts besides characterising impersonators over OSN. The issues found in the literature are largely related to fake news and fake reviews that can influence people to make wrong decisions that would be right decisions to some people who wontedly spread astroturfing reviews. There is no Group-Author model found in the literature that can help in discovering latent social astroturfing groups. In this paper we proposed a methodology for doing this. To our knowledge it is for the first time to have a comprehensive approach to uncover latent astroturfer groups operating to create artificial sentiments over products and services.

PRELIMINARIES AND PROBLEM DEFINITION

This section throws light into details related to various definitions use in this paper and formulation of the problem addressed.

Review: It is an evaluation of a publication or product or service or a movie or company. Author: A person who makes review.

Document: It is a set of sentences. In this paper it is considered interchangeable with review since each review is treated as a document.

Dataset: It is a collection of reviews. In this paper the reviews are related a hotel. **Astroturfing:** It is an organized process to promote or demote a product or service intentionally. It is done by a group of people with biased intentions.

Astroturfing group: It is a group of individuals who put an organized effort to promote or demote a product or service to achieve their intended goal. Generally their review content has high level of similarity and their campaign has certain time window.

Latent Dirichlet Allocation: It is a generative statistical model in Natural Language Processing (NLP) which brings about a set of observations pertaining to latent unobserved things that determine the reason for similarity of some parts.

Problem Formulation

Online reviews assumed unprecedented significance in the contemporary digital world. Reviews play pivotal role in influencing people to make decisions in different domains. People of all walks of life are directly or indirectly impacted by reviews pertaining to products or services. Moreover, the concept of reviews went on to create new business horizons in the real world. Since the impact of reviews is very high on all stakeholders of any business that is associated with products or services on which reviews are made. Literature revealed that reviews are able to either promote reputation of a service or product or making harm to it. Another important fact found in the literature is that reviews may be made intentionally to promote or demote services or products of an organization. It is done by an organized group of people. This phenomenon is known as astroturfing. Astroturfing is extremely hard to detect. It is the challenging problem to be addressed. Therefore detection of astroturfing and astroturfing group is the potential research area which has very high impact on the business world. Creating an author model is essential. Modelling authors and investigating astroturfing author group is aim of the research. The existing research as explored in [1], [2], [3], [6], [8], [9], [10] and [13] focused on various approaches for modelling authors and topics and succeeded in mining text corpus for various applications. Other researchers in [4], [5], [7], [11] and [12] used different approaches such as LDA, variants of LDA and model based clustering. However, there is little research on detection of latent group of social astroturfers.

The above problem is addressed in this paper by proposing a comprehensive methodology. Here, reviews are considered as documents for ease of modelling. Therefore, in this research document and review are interchangeable. A document is denoted as d. Each document is written by an author denoted as a. Here modelling authors is NP-hard. The rationale behind this is that the model should reflect features pertaining to review content, behaviour, and time posted. This model should effectively help to derive the latent astroturfing group behind potential astroturfing. The model should cover both similarity in the content produced by a group of authors and time window. Deriving a threshold for time window to for effective detection of astroturfing group is also a problem to be addressed. A generative model for documents, authors, coupled with time window and machine learning such as unsupervised learning method are required to solve the problem. This paper proposes a methodology with underlying graphical model, algorithms, similarity measures and evaluation metrics for detection of latent astroturfing group. The model used is known as Group-Author (GA) model.

PROPOSED METHODOLOGY

The aim of this methodology is to elaborate the procedure followed to model author-groups from text corpora and detect latent astroturfing group. There are two essential detection problems involved in this methodology. First, it is important to identify astroturfing reviews. Second, it is essential to identify group of individuals behind the identified astroturfing activity. Modelling this kind of problem is NP-hard as it involves a class of problems that are complex in nature besides having high impact on business community across the globe. To understand the problem, model it and solve it we considered a dataset [14] containing thousands of reviews made by different authors. The dataset contains possible distribution of genuine and astroturfing reviews. Modelling the dynamics of such dataset is the challenging problem considered. Especially with respect to Group-Author model that has potential to detect latent astroturfing groups.

Author Model

The author model is meant for modelling authors and documents or reviews. It is a generative model which represents a set of authors and set of reviews. The LDA is not directly used for the reason aforementioned. Instead, the variant of LDA used by Rosen-Zvi [1] is provided in this sub section. We further improve it to make it Group-Author model as discussed in the next sub section. Modelling interests of authors is the purpose of this model. The graphical representation of Author Model is as sown in Figure 1.

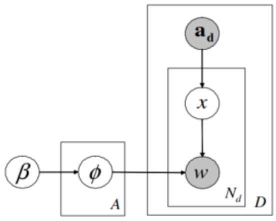

Figure 1: Author model [1]

The boxes are known as plates here. There are many plates in this model. They represent replicates. The left most plate is a replicate of authors. Right side, the outer plate represents set of documents while the inner place represents repeated choice of words within a document. Here x refers to an author of a given word while $\mathbf{a}_d$ indicates a set of users who produced all words. A probability distribution over words is denoted as Ø which is associated with each author. The probability distribution is generated from a symmetric Dirichlet prior denoted as β. Probability distribution has potential to understand author similarity. This author model has limitations. It only provides author information and the words in their documents. It cannot reveal beyond that. This potential limitation is overcome by using Group-Author Model proposed in this paper.

Notations Used in Group-Author Model

Table 1: Notations used in Group-Author model

Notation	Description
A	Set of authors
X	an author of a given word
D	Set of documents representing corpus
α	Author distributions in corpus
β	Astroturfing group author distributions in corpus
θ	Distribution, over words associated with each user
Ø	Distribution, over words associated with users of astroturfing group.
a	A single author
d	A single document from corpus
$\mathbf{a}_d$	Set of users who produced document d.
N_d	Repeated choice of word in give document *d*.

Group-Author Model

It is a generative probabilistic model which yet another variant of LDA. In this model each document is characterized by a distribution over words. No topics are considered. The rationale behind this is that it is derived from Author model which focuses on only authors and their documents rather than underlying topics in the documents. The documents are also represented as random mixtures over latent astroturfers. This is the novelty in this model which is, to our knowledge, proposed for the first time. Therefore this model needs, according to LDA influenced approaches, an outer plate representing documents and inner place representing words.

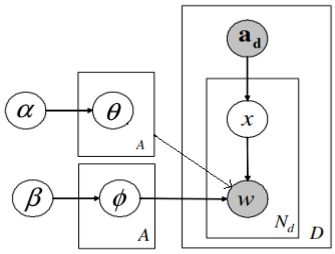

Figure 2: Proposed Group-Author Model (adapted from [1] with two Dirichlet priors)

The Group-Author model has two text corpus level parameters as shown in Figure 2. They are author distributions and latent astroturfing group distribution. They are denoted as α and β respectively. Each author is associated with a distribution over words denoted as θ which is chosen from symmetric Dirichlet (α) prior. In the same fashion, each astroturfing author is associated with distribution over words that reflect astroturfing behavioural indicators denoted as Ø, which is chosen from symmetric Dirichlet (β) prior. By estimating parameters such as θ and Ø, information about which content is written by authors typically and which content written by authors belonging to latent astroturfing groups is known.

There are some important decisions made to make the solution straight forward. The first decision is, not to use LDA model directly as it tends to topics in the text corpora. However, this paper focuses on only author model considering grouping as well. Therefore it is decided to use a variant of LDA proposed by Rosen-Zvi et al. [1]. The second decision is to have machine learning such as unsupervised learning using an effective clustering method. The rationale behind this is that clustering can help in grouping similar reviews and associated authors considering the problem of investigating to discover latent astroturfing groups from text corpora made of hotel reviews.

Comparison with LDA

The original LDA (Figure 3a), author model (Figure 3b) and group-author model (Figure 3c) are compared in this section. The author and group-author models are explained in the previous sub sections.

International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 3 (2018) pp. 1628-1640

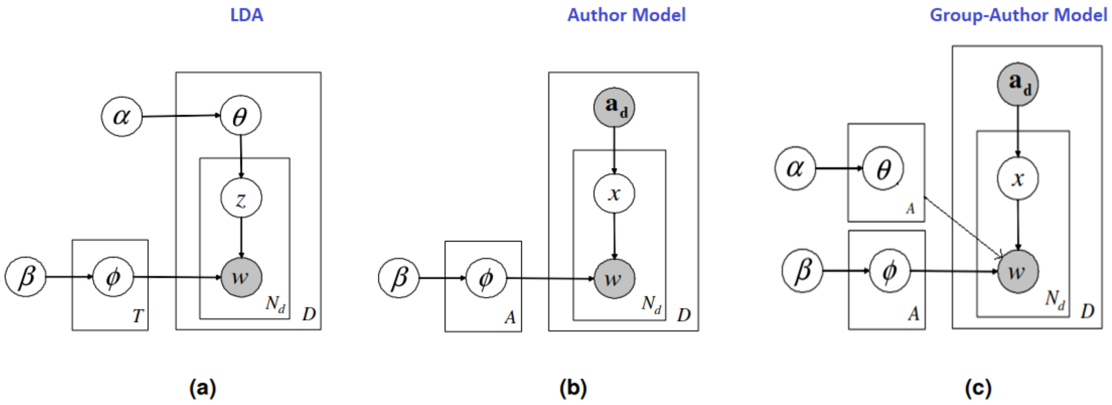

Figure 3: LDA (a), Author Model (b) and Group-Author Model (c)

The LDA model is taken from [1] while the author model and group-author model have sufficient information given in previous sub sections. Where α and β are corpus level parameters. Each author distributed over documents is denoted as θ and it is associated with α. Astroturfing group behavioural indicator is denoted as $\emptyset$ which is associated with β. A set of authors who produced document d is denoted as a_d. Repeated choice of word in given document d is denoted as N_d. Set of authors is denoted as A. Set of documents representing corpus is denoted as D. A word in document is dented as w while x refers to an author of a given word. Based on the work in [36], the following equations provide probabilities involved in the proposed model.

$$p(\theta, \emptyset, X_d, Z_d, W_d \mid \propto, \beta, a_d) = p(\theta \mid \propto) p(\emptyset \mid \beta) \pi_{n=1}^{N_d} p(x_{dn} | a_d) p(w_{dn} \mid \emptyset_{a_n, x_{dn}}, \theta a_d)$$

By getting summary of author distribution and group distributions, marginal distribution of document is obtained as follows.

$$p(X_d, Wd \mid \propto, \beta, a_d) = \iint p(\theta \mid \alpha) p(\phi \mid \beta) \pi_{n=1}^{N_d} \sum_{x d_n} p(x_{dn} \mid a_d) p(w_{dn} \mid \emptyset_{dn,} x_{dn,} \theta a_d) d\emptyset d\theta$$

A product of all distribution probabilities of individual documents and the probability of corpus are obtained as follows.

$$p(D \mid \propto, \beta, a) = \pi_{d=1}^{D} p(w_d \mid \propto, \beta, a_d)$$

Astrotrufing User Group's Behavioural Indicators

With respect to astroturfing and detecting the underlying group of users known as astroturfers, it is essential to consider many astroturfing behaviour indicators as discussed in [15]. The behavioural indicators are time window, content similarity, early review, and ratio of group size.

Time Window

Members who have intention to promote or demote a product or service will form as a group and perform similar activities at the same time. Therefore, temporal domain plays vital role in this research while investigating the behavioural indicators of astroturfers. For a chosen target product, atrosurfers are likely to have the behaviour of working together for some time at least. This active duration in which the members of group make reviews that are not genuine is called time window. In other words, the degree of active involvement of authors or users is known as time window. Each author is denoted as a while group of authors is denoted as g. Therefore Eq. 1 is used to compute time window.

tw(g, p) = latDate(g,p)-earDate(g,p) (1)

where p denotes a product, latDate and earDate denote latest date and early date. The difference between latest data and early date for given group of astroturfers and product is the time window.

Content Similarity

There are two reasons for content similarity. The first reason is that group members copy reviews of other group members. It is known as group content similarity. The second reason is that one or more group members who do not know other members in the group may copy their own previous reviews for similar products. It is known as group member content similarity. It is computed as follows.

CS (g,p) = gcs + gmcs (2)

Early Review Indicator

This is another important behavioural indicator that determines the probability of astroturfing. The rationale behind this is that astroturfers want biggest impact of their activity pertaining to promoting or demoting a product. It is computed as follows.

ER (g,p) = latDate – availDate (3)

Where latDate is the latest date from which astroturfers started their activity in reviewing and availDate is the date from which the product or service is made available.

Ratio of Group Size

The size of astrotufers' group is also another indicator required by Group-Author model proposed in this paper. The ratio of size of group to the total size of reviewers can indicate astroturfing activity. When the group members are the only members making reviews to control sentiments of people on a product, it has more impact. On the other hand, if the genuine reviewers are very large in number, then the astroturfers produce less impact. This ratio is computed as follows.

$$GSR(g,p) = |g|/|G| \tag{4}$$

Where G is the size of all reviewers of the product.

Similarity Measures

Content similarity in the methodology can be found with different similarity functions. They are used to compare two documents or reviews and find similarity. Finding similarity has important utility in this paper as it determines the clustering decisions. Some of the popular functions are as follows.

Jaccard Function

$$\mathrm{sim}(d1, d2) = \mathrm{j}(d1, d2) = f1 \cap f2/f1 \cup f2 \tag{5}$$

Cosine Function

$$\mathrm{sim}(d1,d2)=\mathrm{C}(d1,d2)= f1.f2/||f1||.||f2|| \tag{6}$$

Euclidean Distance

$$sim(d1, d2) = Ec(d1, d2) = \sqrt{(f1 - f2).(f1 - f2)} \tag{7}$$

Extended Jaccard Function

$$sim(d1, d2) = EJ(d1, d2)\frac{f1.f2}{f1.f1+f2.f2-f1.f2} \tag{8}$$

Dice Function

$$sim(d1, d2) = D(d1, d2) = \frac{2f1.f2}{f1.f1+f2.f2} \tag{9}$$

Jensen-Shannon Divergence

Though there are many similarity functions, in this paper, we use the Jensen-Shannon (JS) divergence function to measure similarity between two documents. It is symmetric measure of the similarity of two pairs of distributions. The measure results in 0 if the distributions are identical, otherwise it results in a positive value. It is the average KL divergence as explored in [9]. The divergence of JS is computed as follows.

$$D_{JS} = \frac{1}{2} D_{KL}\ (P\ ||R) + \frac{1}{2} D_{KL}\ (Q\ ||R) \tag{10}$$

$$R = \frac{1}{2}\ (P + Q) \tag{11}$$

The KL is computed as follows.

$$D_{KL}\ (A||B) = \sum_{n=1}^{M} \emptyset_{na} \log\frac{\emptyset_{na}}{\emptyset_{nb}} \tag{12}$$

Latent Astroturfer Group Detection Algorithm

This algorithm is meant for producing initial social astroturfer groups discovered from a corpus of text documents. The inputs and outputs of the algorithm besides its internal process are described in this section. After completion of this algorithm, the resultant groups are subjected to filtering to through the temporal behaviour of astroturfer groups in the real world.

Algorithm: Latent Astroturfer Group Detection (LAGT) Algorithm

Inputs : Review Dataset *D*

Output : Initial Social Astroturfer Groups *G*

Making Document Corpus

1: Initialize document corpus vector *D'*

2: Initialize matrix of vectors *V*

3: Initialize Astroturfer group vector *G*

4: **For** each instance *d* in *D*

5: Extract review into a document *d'*

6: Add *d'* to *D'*

7: **End For**

Pre-processing

8: **For** each document *d'* in *D'*

9: Perform stop words on *d'*

10: Perform stemming on *d'* using Porter Stemming algorithm

11: **End For**

Generating TF/IDF Matrix of Vectors

12: **For** each document *d'* in *D'*

13: **For** each word *w* in *d'*

14: Generate TF/IDF matrix as vector *v*

15: **End For**

16: Add *v* to *V*

17: **End For**

Finding Latent Astroturfer Groups

18: Apply Group-Author model based on K-Means Clustering algorithm

19: Use cosine similarity to group vectors in *V* to form document clusters

20: Group associated authors into social austroturfer groups *G*

21: Return *G*

Algorithm 1: Latent astroturfer group detection algorithm

Algorithm 1 presents the process of discovering latent astroturfer group using text clustering phenomenon. The algorithm takes dataset containing reviews in Excel file format as input. The dataset contains attributes like author, review and date on which review is made by the author. The data present in Excel file is to be subjected to text mining. For convenience, the dataset is converted into a document corpus denoted as D' in the algorithm and denoted as α in the Group-Author model formally. Once document corpus is ready, it is subjected to pre-processing. Pre-processing is made in two phases. In the first phase, the corpus is subjected to stop word removal. Stop words are the words in the set of documents (corpus) containing certain words that do not make any difference in the text clustering process. They are shown in Listing 1.

a,able,about,across,after,all,almost,also,am,among,an,and,any,are,as,at,be,because,been,but,by,can,cannot,could,dear,did,do,does,either,else,ever,every,for,from,get,got,had,has,have,he,her,hers,him,his,how,however,i,if,in,into,is,it,its,just,least,let,like,likely,may,me,might,most,must,my,neither,no,nor,not,of,off,often,on,only,or,other,our,own,rather,said,say,says,she,should,since,so,some,than,that,the,their,them,then,there,these,they,this,tis,to,too,twas,us,wants,was,we,were,what,when,where,which,while,who,whom,why,will,with,would,yet,you,your

Listing 1: Stop words used in the text mining process

After removing stop words, the corpus is ready for processing. However, before processing it is good approach to have stemming process which identifies root words and removes all derived words. The well known class PorterStemmer algorithm is reused here for stemming mechanism. With stemming, the pre-processing ends. Now the documents in corpus are devoid of stop words and derived words. Now the corpus is ready for textual analysis. At this stage TF/IDF matrices are created one for each document. TF/IDF stands for Term Frequency/Inverse Document Frequency. It is a standard measure to reflect importance of a word to a document with respect to corpus. In fact the vectors generated reflecting all documents contain information that can help in clustering process.

While performing clustering, we use Group-Author model where two corpus level parameters are utilized. They are denoted as α and β respectively. The former refers to author distribution in corpus while the latter denotes astroturfing group author distribution in corpus. For grouping purpose, the Group-Author model with K-Means algorithm is implemented. It generates clusters from collection of TF/IDF matrices that reflect astroturfing review clusters. Associated to these clusters, corresponding authors are grouped as per Group-Author model proposed in this paper. The initial social astroturfer groups are identified and reported in the form of a set of clusters denoted as G. Now the G needs to be subjected to temporal dimension of astroturfer group detection process.

Temporal Filtering Algorithm (TFA)

The Group-Author model resulted in the LAGT algorithm contains a set of clusters. Each cluster contains a group of latent social astroturfers based on the content similarity explored in the G-A model. Now the model needs to be updated by validating temporal behaviour of astroturfers in general. For this the author distribution parameter along with the date on which review was given are considered. It is understood that the astroturfers are active in three days period. Based on this assumption the temporal filtering algorithm is defined.

Algorithm: Temporal Filtering Algorithm

Input: Initial social astroturfer groups *G*, temporal threshold *t*

Output: Filtered and temporally validated social astroturfer groups *G'*

```
Initialize astroturfer vector G'
Initialize author vector A of Group-Author model
Initialize date vector D that corresponds to A
Initialize temporal difference vector T
Map A to D to For AD
For each social astroturfer group g in G
  For each author a in g
    Get date d from AD
    Generate temporal differences and populate T
    IF temporal difference satisfies threshold t Then
      Add a to g'
    End IF
  End For
  Add g' to G'
End For
Return G'
```

Algorithm 2: Temporal filtering algorithm

As shown in algorithm 2, the initial social astroturfer groups are subjected to have filtering based on temporal dimension based on the time window explained in the proposed Group-Author model. The authors associated with the documents (reviews) are mapped to corresponding dates. The resultant map is used to obtain data for given author. The dates of users are compared in order to see that all astroturfers who gave reviews in 3 days gap are considered to be true astroturfers. Based on this assumption, the threshold value is compared for each user and decision is made to remove or not to remove user from the group. Once filtering of authors is made from astrotrufer groups, the final list of astroturfer groups is returned by the algorithm. Thus the proposed Group-Author model can be used to have effective text mining and the model can get updated from time to time in order to reflect the latest corpus.

EXPERIMENTAL DESIGN

Experiments are made to evaluate the proposed Group-Author model. The document corpus is obtained from the dataset collected from [14]. The experimental are made with 50 datasets. Each dataset contains 200 instances with attributes such as author name, review (considered as document) and date on which review has been made. Datasets are prepared in such a way that it consists of 100 genuine and 100 fake reviews. This is done for balancing the data and to have controlled experiments for better observations. JDK 1.8 is used to support Java programming language. NetBeans IDE is used to have rapid application development (RAD) features. A prototype application is built using Java programming language. It is a command line application that takes dataset in the form of Excel file and generates astroturfer groups based on the K value provided for the underlying K-Means algorithm used in the proposed Group-Author model. The reviews in dataset are converted document corpus. The application needs stopwords.txt file containing stop words. The Group-Author model is implemented using the parameters of the model as used in the proposed algorithms. The experiments are made in a PC running Windows 10, 64 bit operating system. It has Intel Core i5-421OU CPU with 1.70 GHz processing power and 4.0 GM of main memory. Two metrics are used to evaluate the results. They are time complexity and space complexity. In other words the time taken and main memory consumed to generate initial social astroturfer groups and temporal filtering are recorded.

EXPERIMENTAL RESULTS

Results are obtained in terms of initial astroturfer groups and final filtered latent social asstroturfer groups. The final results are obtained in terms of atroturfing groups, execution time and memory consumed. Results are obtained for 50 datasets (50 restaurants). An excerpt from the results of Barbacco dataset are presented in Listing 2.

```
Astroturfer Group 1: [hhakim, abdulmosimal, kl61, ammarkhaled, Maaxxxdaa, Sandii M, faisalkadadah, Zafer B, Zainal143, Richard G, Inspector G, TPK751, Zain M, PKBhakat, Aathenaa, jimthetravelerr]

Astroturfer Group 2: [jainnemichand, Gordon M, aminfa91, Vishalmiri344, Ash0518, rezask, Richard L, josemanuel04, riicchaaa456, abbas455, AbilioRP, Samar E, Chandra Mohan J, Sana G, anwar19582017, ednaidaesteban, Muhanad K, iaskos]

...

Astroturfer Group 9: [Ibrahim A, Alladsprom, Brenda L., zaid_wasati, J G, jaimecor123, Soundarya C., martinestaban45678, arwafarraj06, J G., Georges Albert , Iram S, Quoc N., aseriahad]
```

Listing 2: An excerpt from end result of Group-Author model

There are nine groups retrieved from the dataset as astroturfing groups. The Group-Author model is applied to the datasets with underlying algorithms. The results pertaining to group statistics, execution time and memory consumed are as follows.

Table 2: Astroturfing group dynamic for different datasets

Dataset Name	Number Of Groups	Count of each Group								
		Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9
Aina	9	24	20	25	21	15	15	14	18	9
Aracely Café	9	16	19	30	25	23	12	25	19	12
Barbacco	9	16	18	31	24	23	31	16	21	14
Beretta	9	16	19	33	22	25	35	28	16	15
BrendasFrenchSoulFood	9	21	19	36	28	23	39	17	28	14
BurmaSuperstar	9	17	21	36	15	23	40	36	17	15
ChaChaCha	9	24	18	36	20	23	40	33	21	15
Coqueta	9	19	18	35	22	23	42	36	23	14
DermRestaurant	9	19	18	35	21	26	49	42	28	15
DumplingKitchen	9	17	20	35	22	23	45	47	29	14
DumplingTime	9	18	19	35	25	23	47	53	23	15
FogHarborFishHouse	9	18	19	36	24	25	53	58	28	14
Frances	9	18	19	37	23	27	52	56	29	15
Francisca	9	25	15	14	31	20	59	63	23	16

GaryDanko	9	20	20	14	33	21	58	68	26	13
Hogwash	9	17	21	38	20	28	25	59	70	14
HoIslandOysterCo	9	17	20	35	29	25	57	62	26	15
HopsAndHominy	9	18	21	15	34	29	21	64	68	26
HRD	9	18	21	15	34	30	21	65	68	27
Hunan	9	18	22	15	34	30	21	66	70	30
IzakayaSozai	9	17	22	14	32	20	91	72	14	12
KElementsBBQ	9	20	22	16	34	31	20	68	74	33
KuiShinBo	9	20	23	14	34	31	21	70	79	37
La Fusion	9	18	23	15	35	31	22	73	79	37
LiholihoYachtClub	9	20	23	14	36	30	20	73	81	37
LittleSkillet	9	21	23	15	39	22	71	83	40	16
Lolo	9	21	23	15	38	22	76	83	37	16
MACD	9	23	24	16	38	30	21	86	88	43
Mano	9	21	25	15	39	21	88	97	44	16
Marlowe	9	23	24	15	37	21	91	102	43	13
MarufukuRamen	9	24	15	37	28	23	87	100	40	26
NojoRamenTavern	9	20	29	16	41	30	23	87	101	50
Nopa	9	23	26	16	39	20	91	102	49	13
ParkerCafe	9	27	26	15	37	22	80	89	45	11
PokiTime	9	23	26	17	39	20	82	104	44	13
QueensLouisianaPoBoy Cafe	9	21	28	17	41	21	80	104	44	16
Ryokos	9	21	29	17	42	30	21	91	109	47
SamWoRestaurant	9	21	29	18	43	23	69	126	59	16
SanTung	9	23	29	20	42	20	102	33	54	13
Skool	9	20	30	20	45	22	100	121	57	12
SottoMare	9	21	29	21	44	20	80	110	61	14
StateBirdProvisions	9	20	32	21	43	30	21	111	139	55
Tacorea	9	21	30	22	44	29	20	108	135	57
TartineBakeryAndCafe	9	21	30	23	44	29	20	108	135	57
TheBeachChaletBrewer yAndRestaurant	9	21	29	22	43	31	23	103	132	59
TheCodmotherFishandC hips	9	21	29	22	43	24	107	133	62	16
TheFrontPorch	9	20	30	23	44	43	24	92	132	77
TheHouse	9	22	31	23	45	23	110	131	68	14
Tropisueno	9	21	31	22	47	25	110	143	65	17
Wayfare Tavern	9	21	31	22	46	23	114	138	69	14

As presented in Table 2, it shows dataset name (restaurant name), number of groups made, and the count of each group formed in the resultant astroturfing groups. All datasets exhibited same number of groups as the datasets contain same number of instances and same k value is given of K-Means algorithm while making text clustering as part of the proposed Group-Author model.

Table 3: Execution time taken for each dataset

Dataset Name	No. of Instances	Execution Time (Milliseconds)
Aina	200	953
Aracely Café	200	656
Barbacco	200	906
Beretta	200	1373
ChaChaCha	200	2144
DumplingKitchen	200	3323
Frances	200	4041
Hogwash	200	4790
HRD	200	5776
KElementsBBQ	200	6979
LiholihoYachtClub	200	7707
MACD	200	9205
MarufukuRamen	200	15765
ParkerCafe	200	13381
Ryokos	200	17250
Skool	200	16958
Tacorea	200	18022
TheCodmotherFishandChips	200	17679
Tropisueno	200	19265
Wayfare Tavern	200	19359

As shown in Table 3, the execution time computed for each dataset while detecting astroturfing groups is recorded. It is measured in milliseconds. The results revealed that each dataset exhibited different execution time requirement. Wayfare Tavern dataset needed 19359 milliseconds for astroturfing group detection while the Aina took only 953 milliseconds time. It took least time while the Wayfare Tavern took highest time.

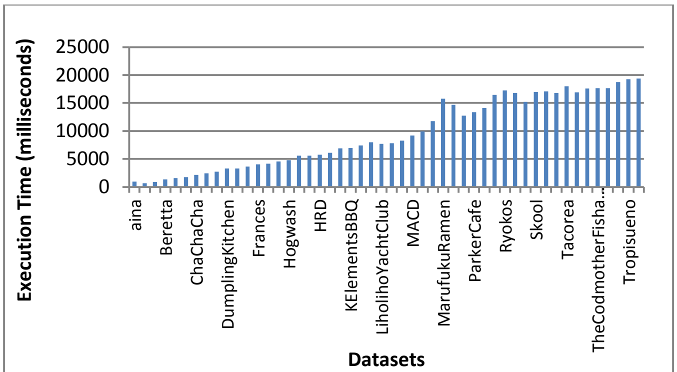

Figure 3: Performance of the proposed system in terms of execution time (milliseconds)

As presented in Figure 3, it is evident that the results reveal different execution time for different datasets though each dataset has same number of instances that is 200. Latent austroturfing group detection took time based on the contents in the reviews. When content is more, it takes more time to compare with other reviews. This is the rationale behind the differences in execution time. The datasets are presented in horizontal axis while the vertical axis represents execution time in milliseconds. The memory consumption dynamics are presented in Table 4.

Table 4: Memory consumed for Astroturfing Group Detection (MB)

Dataset Name	No. of Instances	Memory Consumed (MB)
Aina	200	22.53942108154297
Aracely Café	200	22.53942108154297
Barbacco	200	22.53942108154297
Beretta	200	22.700416564941406
ChaChaCha	200	28.3052978515625
DumplingKitchen	200	28.3052978515625
Frances	200	28.3052978515625
Hogwash	200	28.3052978515625
HRD	200	29.057754516601562
KElementsBBQ	200	33.19554138183594
LiholihoYachtClub	200	41.548805236816406
MACD	200	41.548805236816406
MarufukuRamen	200	44.46788024902344
ParkerCafe	200	55.17707061767578
Ryokos	200	55.17707061767578
Skool	200	63.099525451660156
Tacorea	200	67.59104919433594
TheCodmotherFishandChips	200	67.59104919433594
Tropisueno	200	71.8556900024414
Wayfare Tavern	200	71.8556900024414

Memory consumption differed for each dataset. Though each dataset contains same number of instances, the memory consumption for detecting astroturfing grounds is different. The results revealed that Wayfare Tavern and Tropisueno dataset consumed 71.8556900024414 MB which is the highest while the least memory is consumed for datasets such as Aina, Aracely Cafe, and Barbaco.

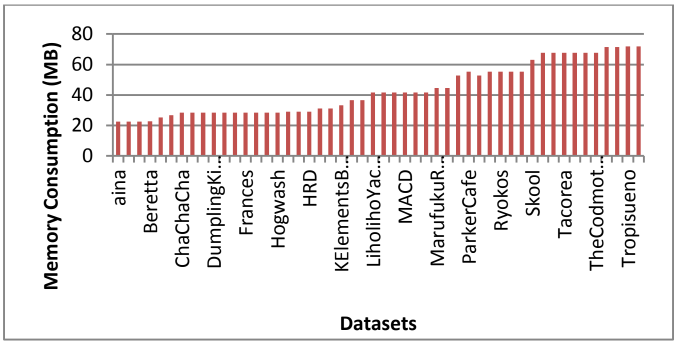

Figure 4: Memory consumed by the algorithms for all datasets (MB)

As presented in Figure 4, datasets are presented in horizontal axis and vertical axis shows memory consumed by the algorithms for different datasets. The results revealed that the datasets consumed differently though they have similar number of instances. Since it is text mining process and similarity comparison among different reviews (small and big) it causes memory consumption differently.

EVALUATION OF PROPOSED ALGORITHMS

The proposed Group-Author model is evaluated with a prototype application. The methodology for evaluation is described here. From software industry 10 people who were aware of data mining and text mining were selected to evaluate the proposed system. The details of the experts participated in the evaluation are kept confidential as part of ethical consideration. The application and datasets are provided to human experts to evaluate the proposed system. They made careful observations on the inputs and outputs and prepared possible astroturfing behaviour and astroturfing groups. In fact, they came up with a ground truth table which helps in performance evaluation. The ground truth values are them compared with the results of experiments made with the system. The results revealed the efficiency of the proposed system. The Table 5 shows confusion matrix for the prepossessed algorithm. We used two statistical measures for evaluation. They are known as precision and recall. The equations (13) and (14) reflect precision and recall respectively.

$$\text{Precision} = (TP/(TP+FP)*100 \quad (13)$$

$$\text{Recall} = (TP/(TP+FP))*100 \quad (14)$$

The precision indicates the ratio of number of correctly grouped reviews to the total number of reviews. Similarly, recall is the ratio that indicates the number of correctly grouped reviews to the total number of correctly matching reviews present in the dataset. Both measures reflect results that are inversely proportional. If one increases, the other decreases. The results of evaluation are as presented in Figure 5.

Table 6: Precision and recall results

	Precision	Recall
LAGD Algorithm	98	12.43

As shown in Table 6, the results revel that high precision is observed besides low recall. The LAGD algorithm is evaluated and found that it have good performance in making astroturfing groups.

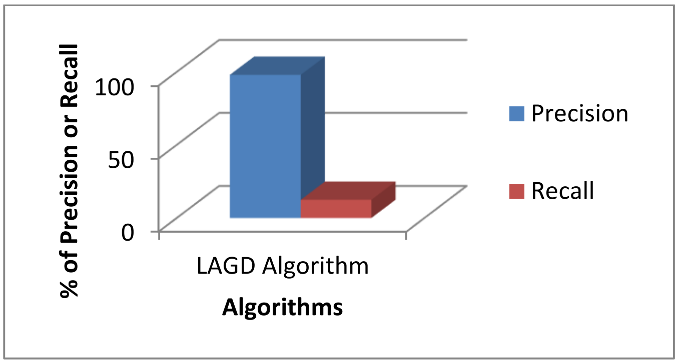

Figure 5: Precision and recall measures

As shown in Figure 5, it is evident that the proposed Group-Author model with underlying algorithm LAGD showed high precision. It reflects the performance of the proposed system as evaluated with ground truth and as per the confusion matrix presented in Table 5.

THREATS TO VALIDITY

This paper proposed a methodology with Group-Author model for latent social astroturfing group detection. The methodology includes two algorithms namely Latent Astroturfer Group Detection (LAGT) and Temporal Filtering Algorithm (TFA) for achieving this. Experiments are made with a prototype application using 50 datasets collected from YELP.COM which facilitates online reviews. The application is able to detect astroturfing groups which are latent in the corpus by using Group-Author model for discovering author and group distributions. The results are evaluated with group truth provided by experts and confusion matrix presented in Table 5. As far as astroturfing and astroturfing group detection is concerned, there are certain threats to validity of the methodology. The first validity concern is the ground truth received from human experts. The rationale is that limited number of people is involved in the evaluation. There might be human errors that went undiscovered. Biased ground truth is another validity threat in such evaluation methodology where human experts are involved. Yet another validity threat comes from the fact that we have used only 50 datasets each one containing 200 reviews. By any means the total of 10000 reviews cannot be claimed as sufficient corpus to generalize the findings. Thus the correctness of ground truth, and limitations in the input corpus are significant threats to validity.

CONCLUSIONS AND FUTURE WORK

Online reviews about products and services have their influence on decision making of people. It is ascertained from the literature that online review web sites do not have adequate measures to control astroturfing reviews. These are the reviews made by organized set of people in certain duration to promote or demote a product or service. In other words reviews made intentionally to influence people are known as astroturfing reviews that cause apprehensions in the mind of people. This is not a good sign of healthy presence of social networking sites where people give reviews or micro reviews. There is no problem with genuine reviews provided by legitimate customers or users of services. The malpractice of some people or organizations pollutes review datasets with fake opinions. Such reviews cause damage when they are used to make strategies or decisions. The existing literature has Author, Top and Author-Topic models for processing textual data using LDA. There is no model that reflects astroturfing authors, astroturfing reviews and astroturfing groups. In this paper we proposed Group-Author model which is based on a variant of LDA. The contribution of this paper is two-fold. First, it detects astroturfing groups from corpus using the proposed Latent Astroturfing Group Detection algorithm and the second one is to validate and filter the group members with temporal criterion. We built a prototype application to demonstrate proof of the concept. The empirical results revealed the utility of the proposed model in terms of discovering astroturfing groups besides reducing time and space complexity. We believe that the proposed Group-Author model can be extended further to have Group-Topic-Author model. Therefore it is an interesting direction for future work.

REFERENCES

[1] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers and Padhraic Smyth. (2003). The Author-Topic Model for Authors and Documents. IEEE, p1-8.

[2]. Dou Shen, Jian-Tao Sun , Qiang Yang and Zheng Chen. (2006). Latent Friend Mining from Blog Data. IEEE, p1-10.

[3]. Michal Rosen-Zvi ,Chaitanya Chemudugunta,Thomas Griffiths and Padhraic Smyth and Mark Steyvers. (2010). Learning Author-Topic Models from Text Corpora. ACM. 28 (1), p1-38.

[4]. Yan Liu, Alexandru Niculescu-Mizil and Wojciech Gryc. (2009). Topic-Link LDA: Joint Models of Topic and Author Community. Machine Learning, p1-8.

[5]. Melnykov and Maitra. (2010). Model-based clustering. IEEE, p1-15.

[6]. Nazneen Fatema, Rajani Kate McArdle and Jason Baldridge. (2014). Extracting Topics Based on Authors, Recipients and Content in Microblogs. ACM, p1-4.

[7]. Wenjia Niu , Gang Li , Hui Tang , Xu Zhou and Zhongzhi Shi. (2011). CARSA: A context-aware reasoning-based service agent model for AI planning of web service composition. Network and Computer Applications, p1-14.

[8]. Diane J. Hu. (2007). Latent Dirichlet Allocation for Text, Images, and Music. IEEE, p1-19.

[9]. Liangjie Hong and Brian D. Davison. (2010). Empirical Study of Topic Modelling in Twitter. ACM, p1-9.

[10]. Johan Risch. (2016). Detecting Twitter topics using Latent Dirichlet Allocation. IEEE, p1-48.

[11]. David M. Blei,Andrew Y. Ng and Michael I. Jordan. (2003). Latent Dirichlet Allocation. Machine Learning Research, p1-30.

[12]. David M. Blei,Andrew Y. Ng and Michael I. Jordan. (2001). Latent Dirichlet Allocation. IEEE, p1-8.

[13]. J.Kisthiya. (2016). Review Selection Using Micro-Reviews. Applications in Information and Communication Engineering. 2 , p1-5.

[14] Hotel Reviews Dataset. Retrieved from https://www.tripadvisor.com.au/Hotel_Review-g295424-d2407420-Reviews-Millennium_Plaza_Hotel_Dubai-Dubai_Emirate_of_Dubai.html#rd_reviews_section_start

[15] Arjun Mukherjee, Bing Liu, and Natalie Glance (2012). Spotting Fake Reviewer Groups in Consumer Reviews. International World Wide Web Conference, France, p1-10.

[16] Leman Akoglu,Rishi Chandy and Christos Faloutsos. (2013). Opinion Fraud Detection in Online Reviews by Network Effects. *International AAAI Conference on Weblogs and Social Media*, p1-10.

[17] Michael Fire,Roy Goldschmidt, and Yuval Elovici. (2014). Online Social Networks: Threats and Solutions. *IEEE Communication Surveys & Tutorials*. 16 (4), p1-18.

[18] Arjun Mukherjee, Vivek Venkataraman, Bing Liu and Natalie Glance. (2013). What Yelp Fake Review Filter Might Be Doing. *International AAAI Conference on Weblogs and Social Media*, p1-10.

[19] Justin Malbon. (2013). Taking Fake Online Consumer Reviews Seriously. *Springer Science+Business Media New York*, p1-22.

[20] Miriam J. Metzger, Andrew J. Flanagin, & Ryan B. Medders. (2010). Social and Heuristic Approaches to Credibility Evaluation Online. *Journal of Communication*. 60 , p413–439.

[21] Michael Luca,Georgios Zervas. (2015). Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. *ACM*, p1-36.

[22] Arash Molavi Kakhki,Chloe Kliman-Silver and Alan Mislove. (2013). Iolaus: Securing Online Content Rating Systems. *ACM*, p1-10.

[23] Romina Cachia, Ramón Compañó, Olivier Da Costa. (2007). Grasping the potential of online social networks for foresight. *Technological Forecasting & Social Change*, p1179–1203.

[24] Saeedreza Shehnepoor, Mostafa Salehi, Reza Farahbakhsh, Noel Crespi. (2017). NetSpam: a Network-based Spam Detection Framework for Reviews in Online Social Media. *IEEE*, p1-10.

[25] Arjun Mukherjee, Vivek Venkataraman, Bing Liu, Natalie Glance. (2013). Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews. *University of Illinois at Chicago*, p1-10.

[26] Michael Anderson,Jeremy Magruder. (2011). Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database1. *ACM*, p1-20.

[27] Chi Zhang , Jinyuan Sun,Xiaoyan Zhu,Yuguang Fang. (2010). Privacy and Security for Online Social Networks: Challenges and Opportunities. *IEEE Network*, p1-6.

[28] Cliff Lampe, Nicole Ellison, Charles Steinfield. (2010). A Familiar Face(book): Profile Elements as Signals in an Online Social Network. *ACM*, p1-10.

[29] Michael Sirivianos,Kyungbaek Kim and Xiaowei Yang. (2010). FaceTrust: Assessing the Credibility of Online Personas via Social Networks. *IEEE*, p1-6.

[30] Bello-Orgaz, G., Jung, J.J. and Camacho, D., 2016. Social big data: Recent achievements and new challenges. *Information Fusion*, *28*, pp.45-59.

[31] George Pallis, Demetrios Zeinalipour-Yazti, and Marios D. Dikaiakos. (2011). Online Social Networks: Status and Trends. *Springer* , p213–234.

[32] Xujuan Zhou, Yue Xu, Yuefeng Li, Audun Josang and Clive Cox. (2012). The State-of-the-Art in Personalized Recommender Systems for Social Networking. *Springer* , p1-15.

[33] C.L. Lai, K.Q. Xu, Raymond Y.K. Lau,Yuefeng Li and Dawei Song. (2010). High-Order Concept Associations Mining and Inferential Language Modeling for Online Review Spam Detection. *IEEE*, p1-8.

[34] Susan Athey and Guido W. Imbens. (2017). The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives*. 31 (2), p3–32.

[35] Oana Goga. (2015). Matching User Accounts Across Online Social Networks: Methods and Applications. *HAL*, p1-151.