Journal of Web Engineering, Vol. 16, No.5&6 (2017) 505-523

IDENTIFYING THE INFLUENTIAL BLOGGERS: A MODULAR APPROACH BASED ON SENTIMENT ANALYSIS

UMAR ISHFAQ

COMSATS Institute of Information Technology, Attock
umer.bravo@gmail.com

HIKMAT ULLAH KHAN*

COMSATS Institute of Information Technology, Wah
**Corresponding author Hikmat.ullah@ciitwah.edu.pk*

KHALID IQBAL

COMSATS Institute of Information Technology, Attock
khalidiqbal@ciit-attock.edu.pk

Received July 6, 2016
Revised March 13, 2017

The social web provides an easy and quick medium for public communication and online social interactions. In the web log, short as a blog, the bloggers share their views in the form of creating and commenting on blog posts. The bloggers who influence other users in a blogging community are known as the influential bloggers. Identification of such influential bloggers has vast applications in advertising, online marketing and e-commerce. This paper investigates the problem of identifying influential bloggers and presents a model which consists of two modules: Activity and Recognition. The activity module takes into account a blogger's activity and recognition module measures a blogger's influence in his/her social community. The integration of activity and recognition modules identifies the active as well as influential bloggers. The proposed model, MIBSA (Model to find Influential Bloggers using Sentiment Analysis), takes into account the existing and novel features of sentiment expressed in content generated by a blogger. The model is evaluated against the existing standard models using the real world blogging data. The results confirm that sentiment expressed in blog content plays an important role in measuring a blogger's influence and should be considered as a feature for finding the top influential bloggers in the blogosphere.

Key words: Social web, Blog, Blogosphere, Influential bloggers, Big Data, Sentiment Analysis
Communicated by: B. White & M. Bielikova

1 Introduction

The creation and usage of the web content have changed a lot over during the last decade with the emergence and evolution of social networking platforms. The online social networks such as Facebook, Google+, blogs etc., have widened the scope of social interaction and allowed its users to generate large volumes of data in a matter of minutes and seconds [1]. The focus on social networks, online user

activities and user generated content brings the scope of the study of users' influence in virtual communities. A blog is an online social platform for bloggers to express their views and opinions in the form of blog posts or comments. The comments on the blog posts are displayed in reverse chronological order with the most recent comments on top of the discussion thread. Blogs contain multimedia content and offer social networking services by creating online communities and provide links to such communities. These online communities form a virtual universe of blogs known as the Blogosphere. Blogs can be classified into two main categories: single-author blogs and community blogs. The single-authored blogs are more like personal accounts or personal diaries whereas the community blogs are more like discussion forums offering a high degree of collaboration to the community members.

In a physical community, people consult their friends or experts before making decisions like buying a product, selecting a place to shop or a movie to watch. The virtual community is quite similar to a physical community where individuals consult others (e.g., friends or experts) on a variety of social issues such as online shopping [1], education and careers [2] or starting a business [3]. These others being consulted are known as the *influential bloggers*. Identification of influential bloggers is a significant phenomenon in blogging community. The influential bloggers use online platforms to spread their influence. Being part of a large online community, the influential bloggers have the ability to sway opinions of community members [4, 5]. The influential bloggers can also render valuable services in online marketing and advertising [6, 7]. For instance, commercial organizations turn influential bloggers into their zealous supporters by winning their trust and save hundreds of millions of advertising expenses [8]. Viral marketing is an important application of influential bloggers. Influential bloggers use social networks to increase brand awareness, therefore, can help companies achieve additional marketing objectives [9].

Various models and techniques have been proposed to identify the influential bloggers. However, the existing research methods lack to consider the sentiment expressed in the blog posts of a blogger. The blog posts reflect the views and opinions of a blog writer. The sentiment analysis helps determine a writer's attitude about a topic. The opinions of a blog writer are investigated to analyse whether the views are positive, neutral or negative. Therefore, we posit that sentiment expressed in a blog post is an important feature and is helpful to identify the influential bloggers. The work of Akritidis at el., [10] is considered as a standard model in the relevant literature and has been considered as a baseline. The proposed model, MIBSA (Model for Influential bloggers using Sentiment Analysis), is based on a modular approach introducing the modules of activity and recognition. The activity module measures a blogger's productivity and the recognition module measures a blogger's reputation and impact in an online social community. With module integration, we aim to find active and influential bloggers in a blogging community. The empirical study is evaluated using the performance evaluation measures of OSim [11], Kendall and Spearman's correlations by applying on the dataset of a real world blog.

The remaining part of the paper is divided into the following sections: section 2 reviews the most relevant literature, section 3 presents the statement of the problem, section 4 presents the proposed framework, section 5 outlines the performance evaluation measures, section 6 discusses the evaluation results and presents comparative analysis with the existing models. Finally, section 7 presents the conclusion and future research directions.

2 Related Work

The research domain of measuring users' impact on others starts from the study of the concepts of prestige, position and prominent roles of the social actors on social media platforms. The literature on identifying the influential social actors can be classified into two main categories: feature based models and social network based models.

2.1. Feature based Models

The literature review presents the most research studies targeting to identify the influential bloggers. The pioneer model of iIndex [12] provides a valuable insight into various blogging related phenomena and outlines an effective strategy to clearly differentiate influential bloggers from active bloggers. The iIndex is based on four novel features from the blogosphere: recognition, novelty, eloquence and activity. The *recognition* determines the reputation of a blogger in the blogging community through the number of inlinks and comments on the blog posts of a blogger. *Novelty* determines the innovative ideas expressed in the blog posts of a blogger. Similarly, *eloquence* of a blogger is determined by measuring the length of blog post contents. Finally, the *activity* is based on the number of blog post initiated by the blog posts of a blogger. The extended version of iIndex, known as iFinder [7], is based on a matrix based approach of PageRank [13]. However, a study suggests that PageRank is not an appropriate choice for ranking blogs as the blogosphere provides a sparse graph [14].

One of the challenges in the blogosphere is identifying the influential bloggers from individual blogs [15]. Majority of bloggers has limited links and are locally connected. Therefore, it is important to extract valuable information from individual blogs as well. Using a clustering approach, the researchers synthesize individual blogs into blogging communities to identify the influential social actors. Temporal features of blog posts are another important aspect of the blogosphere. Due to the dynamic nature of blogs, blog posts get old within a short span of time and, therefore, lose influence. Akritidis et al. [16] takes into account the link structure of blogging networks and the temporal features of blog posts to identify the current or most recent blogger influence. However, the work fails to clearly discriminate blogger influence from blogger productivity in the blogging community. The indices of BI and BP [10] effectively separates the blogger influence from blogger productivity in an online community. The productive bloggers write blog posts on a regular basis, whereas influential bloggers influence other online users. It considers the temporal influence of a blogger, therefore, the models identify the recent influence of a blogger. However, it does not introduce a feature. Integrating product and the user information with the temporal aspects of the user reviews on products can improve the sentiment classification of the reviews [17]. However, this sequence modelling using neural networks performs the sentiment analysis of standard documents only. The sentiment mining systems require more advanced brain inspired reasoning methods that are psychologically motivated with a comprehensive knowledge base of common sense [18, 19]. A recent work [20] studies sentiment from opinion videos via gestures and spoken words with reduced speaker bias. However, the proposed multimodal dictionary for predicting performance needs a profound knowledge base.

SIIB (Semantically Identified Influential Bloggers) [21] is a mining algorithm that uses the influence factors like content semantics of blog posts of bloggers, quantitative analysis of the contents and fellow readerships with their comments on the blog posts to identify the influential bloggers. The work lacks to compare its results with existing models. On the other hand, I-FBCount (Influence-FacebookCount) [22]

introduces the features of uniqueness and Facebook count to measure blogger originality expressed in the blog posts and the importance of emerging social media platforms respectively. The model distinctively ranks the influential bloggers by combining novel features with link structure of the blogging platforms. However, I-FBCount relies mainly on popularity measures like g+ and Facebook to identify the influential bloggers. PInf (Post Influence) [23] is an effective algorithm which measures the quality of blog posts by evaluating comments and shows higher accuracy by excluding the self-comments of bloggers in their own blogging threads. On the other hand, the Alghobiri et al. [24] explores the correlation of sentiments and top bloggers. However, Khan et al. [25] proposed modular approach based on a comprehensive set of features of bloggers and blog posts including a novel feature from the blogosphere.

2.2. Graph based Models

The link based or network based models use the properties of social media platforms. For instance, Li et al. [27, 28] proposed pure marketing oriented neural network models which combine the network and content based features of blog posts with activeness based features of a blogger. Network features determine the range of blogger influence in the social community, whereas content based features reflect a blogger's view on marketing a certain product. Similarly, the activeness features take into account a blogger's willingness to interact socially with other members in the social community. The drawback of neural network model is that it is based on the well-known page ranking algorithm and heuristics which are not suitable for highly dynamic blogging platforms [14]. Goyal et al. [29] investigates the influence propagation in social networks by introducing algorithmic based probabilistic causation for influence maximization in online communities. The influence propagation is based on the novel concepts of "Leaders" and "Tribal Leaders". However, the algorithm does not clearly differentiate the influence propagation of the leaders and the tribal leaders. Kayes et al. [30] applies to aggregated centrality approach and exploits the link structure of the social networks to identify power users. On blogs, the nodes having information on a topic of interest are known as "authorities". Similarly, a hub points to the best authorities in the blogging community.

As the influential bloggers are strongly connected and form the core of the blogging networks, therefore, the nodes with higher authority are pointed to by many hubs. Bui et al. [31] applies h-index and its variants and measures blogger productivity in the social community. An important aspect of h-index is that it captures the overall influence by taking into account the influential and non-influential blog posts. On the other hand, Qureshi et al. [14] proposes a clustering based approach that group blogs sharing a common interest in a cluster. The clustering is based on the topic of blog post contents and clusters are ranked according to the level of interest shown towards the topic. However, the TDIR lacks adequate experiments to identify the topic clusters. Also, the dataset is not large enough to validate the model. Shang et al. [32] proposed time efficient node centrality (sub-modularity) framework for influence maximization on large scale social networks. The framework utilizes a weighted cascade approach to integrating the seed expansion as well as the influence propagation within different, independent virtual communities to compute a sub-modular influence score. Gong et al. [33] proposed a memetic algorithm based on multi-level learning approach to optimize the 2-hop influence spread. The model detects and clusters the significant communities and find the ultimate seeds from the candidate pool of the nodes. However, the dataset used is small and does not incorporate sufficient properties of social networks. As a social network connects millions of nodes, therefore, it becomes hard to understand

and visualize nodes and edges. Smith et al. [34] designed a pruning technique applicable on social networks by retaining a subset of the original network, which, in turn, presents a refined outlook of the entire social network. Calderoni et al. [35] applies the community detection and analysis methods to study the clustered structure of mafia networks and explores the co-participation and role of individuals in criminal organizations and rendezvous. Erik Cambria proposed an emotion processing sentiment mining technique for automated up keeping of social and product reviews which results in improved customer relations and enhanced recommendations.

3 Problem Statement

For a given set *U* of *M* bloggers, U = $\{u_1, u_2, u_3, u_4, \ldots\ldots\ldots u_M\}$, the problem of identification of the top influential weblog users is formally defined as finding an ordered subset V of N bloggers ordered as per their respective scores of influence, S_{inf}, such that V ⊆ U and N ≤ M, i.e., $S_{inf}\ (u_{j1}) \geq S_{inf}\ (u_{j2}) \geq \ldots\ldots\ldots S_{inf}\ (u_{jN})$. The set V consists of N top influential users of the weblog. The influence score S_{inf} is based on the modules of blogger activity (S^b_{act}) and recognition (S^b_{rec}) on a blog site.

4 The MIBSA Model

The proposed framework, MIBSA (Model to find Influential Bloggers using Sentiment Analysis) is proposed to identify the top influential bloggers. The proposed framework, shown in Figure 1, based on the following: (1) The modules of Activity and Recognition (2) The features of the modules (3) Computation of blogger activity and recognition in the blogging community and (4) The integration of the modules to identify the top active and influential bloggers. We took a modular approach and present the modules of activity and recognition. The activity module identifies the active bloggers; the bloggers not only consistent in writing blog posts but also present objective and technical information on various topics being discussed online.

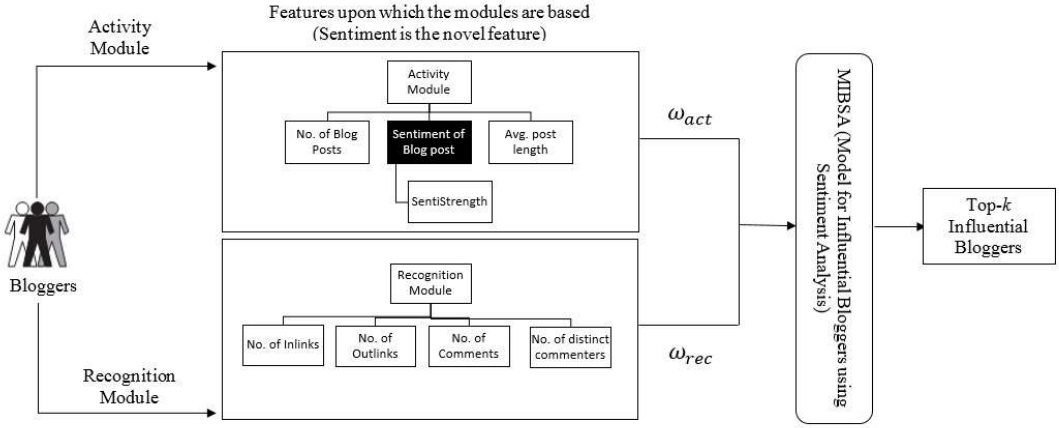

Figure 1 The Framework of the proposed MIBSA Model.

On the other hand, the recognition module identifies the bloggers, well-reputed and recognized in their blogging communities. The blogger activeness is based on (1) Number of blog posts of a blogger (2) The average length of blog posts and (3) Sentiment Analysis of blog posts. Similarly, blogger

recognition is based on (1) Number of inlinks to a blog post (2) Number of outlinks to other blogsites (3) Number of comments on a blog post and (4) Number of distinct commenters. The blogger features are extracted from the real world blog application i.e. Engadget[a] dataset. First, we compute the numerical score for blogger activity and blogger recognition and subsequently, we integrate the activity recognition scores to identify active as well as influential bloggers in a social community.

Table 1 Notations used in MIBSA Modules

Notation	Meaning
S_{act}^{b}	Activity score of a blogger b
W_{avglen}	Weight of average post's length
N_{posts}	Number of blog posts of a blogger b
S_{pos}^{b}	Positive sentiment score of a blogger b
S_{neg}^{b}	Negative sentiment score of a blogger b
S_{obj}^{b}	Objective sentiment score of blogger b
S_{rec}^{b}	Recognition of a blogger in the community
$W_{v.com}$	Weight of unique commenter [0..1]
N_{com}^{b}	Number of comments to the blog posts of a blogger b
$N_{inlinks}^{b}$	The number of inlinks/citations to the blog posts of a blogger b
$N_{outlinks}^{b}$	Number of outlinks from the blog posts of a blogger b to other blogs

In figure 1, ω_{act} and ω_{rec} represents the weights for activity and recognition modules respectively which represents the importance of each module in finding the optimal ranking of top influential bloggers.

4.1. Activity Module

A blogger who initiates blog posts on a regular basis is considered as an active blogger. The activity module measures a blogger's activeness and productivity in the social community. The module is based on the following features:

4.1.1. Blog posts

The number of blog posts represents blogger activeness in the social community. An active blogger consistently writes the blog posts.

4.1.2 Sentiment Feature

Sentiment analysis aims to determine a writer's attitude with reference to some topic or a document's general contextual polarity [37, 38]. In sentiment analysis, the basis task is the classification of polarity of the text [39]. In classification, a writer's opinion is analysed to see whether it is positive, negative or objective [40]. In this paper, we quantitatively analyse the contents of a blog post using one of the standard tool SentiStrength[b] which is based on principles of opinion mining. SentiStrength automatically

[a] https://www.engadget.com/tag/liveblog/
[b] http://sentistrength.wlv.ac.uk/

performs sentiment analysis of up to 16,000 texts on social web with human level accuracy. It also allows easy inclusion of five other languages besides English. SentiStrength uses two reporting mechanisms given as under:

- -1 to -5 (not negative to extremely negative)
- +1 to +5 (not positive to extremely positive)

The reason behind two reporting mechanisms is based on a theory in Psychology that humans process both emotions, positive and negative, at the same time [41].

For sentiment computation in SentiStrength, the input file is a list of texts, one per line and Output file is a copy of the texts, plus the classifications. SentiStrength can process not only one text, but also multiple texts. SentiStrength uses the opinion mining technique to compute the sentiment from multiple texts by classifying the sentiment of each line of the file separately. For instance, for the following multiple texts:

I just thought that I would say HI... ----- Love you After the series it looked like shit!! Damn its been a good while that i don't see u.

The SentiStrength classifies the text as:

4 1 I just thought that I would say HI... ----- Love you

1 4 After the series it looked like shit!!

3 2 Damn its been a good while that i don't see u

The optimization step alters the sentiment dictionary term weights to fit the data better, e.g., love (+4) -> love (+3). SentiStrength can be modified for new languages or domains. It needs linguistic work, not programming work, to modify.

4.1.3 Average length of a blog post

The average length of a blog post is taken as the eloquence measure. The logic behind the eloquence is that there is no real incentive for writing longer pieces of text, therefore, the users sharing lengthy blog posts have real incentive to present valuable information. Average post length has a positive association with number of received comments i.e. lengthy blog posts normally get more attention from users of social media and receive more comments.

The blogger activity is computed using equation (1) given as under:

$$S^b_{act} = W_{avglen} * N_{posts} + \left[\left(S^b_{pos} - S^b_{neg}\right) + S^b_{obj}\right] \qquad (1)$$

In real life, people are influenced by facts and figures and the positive content. In this study, we follow a real life scenario. Therefore, negative sentiment (S^b_{neg}) is reduced from the positive sentiment (S^b_{pos}) to find the overall positivity and next, we add the objective sentiment (S^b_{obj}) which represents the facts and figures. In equation (1), W_{avglen} acts as the weight factor i.e. the number of blog posts N_{posts} is multiplied by the length of the blog posts W_{avglen}.

4.2. Recognition Module

Recognition represents the reputation (or authority) of a blogger within the blogging community. Recognition measures the blogger influence and is based on the following features:

4.2.1. Inlinks

While doing research, a researcher cites the other researcher's work to prove his/her research findings. The author being referenced is regarded and respected by the research community. Similarly, the inlinks are like citations to the blog posts of a blogger. The feature implies direct authority and the influence of a blogger in the social community. A blogger whose blog posts are receiving a number of inlinks is considered highly influential.

4.2.2. Outlinks

Outlinks imply lesser novelty in blog posts. Presence of outlinks decreases originality of the ideas as the blogger is referring to the contents of someone else in the social community.

4.2.3. Comments

Comments are an important part of blogger recognition. Higher number of comments represent the likeness and the popularity of the blogger in the social community.

4.2.4. Unique commenters

Unique commenters are online users inspired by the blog posts initiated by influential bloggers and write comments. We introduce a feature to measure how unique commenters are being influenced by the blogger. It is a good measure as merely measuring the comments may not only be proper as few bloggers can add too many comments so this additional factor is direct feature to count the number of influenced bloggers. The number of distinct commenters is directly proportional to the degree of a blogger's influence. Together, the number of inlinks, comments and distinct commenters determine the true recognition of a blogger in the social community.

The blogger recognition is computed using equation (2) given as:

$$S_{rec}^{b} = W_{v.com} * \frac{N_{com}^{b}}{N_{posts}^{b}} + \left(N_{inlinks}^{b} - N_{outlinks}^{b}\right) \qquad (2)$$

Outlinks $N_{outlinks}^{b}$ imply lesser novelty, therefore, it is reduced from the inlinks $N_{inlinks}^{b}$. In equation (2), number of comments N_{com}^{b} is divided by number of blog posts N_{posts}^{b} of a blogger to normalize the values in the range [0,1].

5 Experimental Setup

The experimental setup consists of three sub-sections. Section 5.1 introduces the data set used for the experimental setup, section 5.3 presents the performance evaluation measures for comparative analysis with the baseline and in section 6, we perform five different types of comparisons. From these comparisons, we try to understand the behaviour of the MIBSA model from different perspectives and to establish logical grounds for concluding and pave the way for future research directions.

5.1. Engadget Dataset

Engadget[c] is a blog covering the news on electronic gadgets in six different languages. It is ranked among the top five blogs of 2010 by the Time magazine[d]. The dataset is widely used in existing relevant literature [10, 22, 41]. Currently, Engadget is operating ten blogs, out of which four are in English. Engadget has six versions of blogging services operating under independent editorial teams. Engadget dataset is active and constantly being used by its users. Table 2 presents the characteristics of Engadget dataset.

Table 2 Dataset Characteristics

Characteristics	**Engadget**
Number of bloggers	93
Number of blog posts	63,358
Number of inlinks	3,19,880
Number of comments	3,672,819

5.2. Baseline

BP and BI index [10] represent a blogger's productivity and blogger's influence in the blogging community and this model is taken as the baseline of the proposed framework. The model considers (only) the recent influence of a blogger by taking into account the number of recent blog posts of a blogger. On the other hand, blogger influence is based on the features of inlinks and comments received on the blog posts of a blogger. However, the model excludes the outlinks in computing blogger's influence. We posit that outlinks represent lesser novelty, therefore, outlinks need to be included as well as reduced from inlinks and comments as these corresponds to direct blogger influence. Moreover, computing only the recent influence utterly ignores the influence of older blog posts of a blogger. The proposed MIBSA model considers a comprehensive approach by taking into account multiple aspects of blogger and blog posts to computer refined value of blogger influence.

5.3. Performance Evaluation Measures

Unlike earlier work, the proposed framework presents a comprehensive comparative analysis using performance evaluation measures given as:

5.3.1. OSim

OSim [11] measures the degree of similarity and overlap between the two ordered lists x and y each of size k. It is mathematically expressed in equation (3) given as:

$$OSim(x, y) = \frac{|x \cap y|}{k} \qquad (3)$$

OSim is used to find out the number of common bloggers between the proposed model (MIBSA) and the baseline.

[c] https://www.engadget.com/tag/liveblog/
[d] http://content.time.com/time/specials/packages/article/0,28804,1999770_1999761_1999863,00.html

5.3.2. Spearman Correlation

Spearman correlation measures the statistical relationship dependence between two ordered lists (variables). The nature of the relationship between the variables is assessed using a monotonic function which preserves the order of the input data. It is mathematically expressed in equation (4) given as:

$$\rho = 1 - \frac{6\Sigma d_{i^2}}{k(k^2 - 1)} \qquad (4)$$

Where d_i represents the difference between ordered lists and k is the size of input data.

5.3.3. Kendall Correlation

Kendall correlation is a non-parametric and distribution free test of independence, which measures the association or dependence between two ordered lists (variables). It represents the variance analysis through which we can easily examine the ranking differences between the ordered lists. It is computed using equation (5) given as:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \qquad (5)$$

Where n_c represents the concordant pairs and n_d denotes the discordant pairs respectively.

5.3.4. Difference between Spearman and Kendall Correlation

Kendall correlation, actually, is the difference between concordant and discordant pairs divided by the sum of concordant and discordant pairs, whereas the Spearman correlation is the difference between the rank orders. Kendall's tau is representing the proportion of concordant pairs relative to discordant pairs, whereas Spearman's correlation detects the rare and unusual sensitives that are very big discrepancies. Kendall's tau has a more intuitive interpretation: number of concordant pairs minus the number of discordant pairs. It outputs better estimates of the corresponding population parameters, especially in smaller sample sizes. It shows higher accuracy when the samples are smaller. However, Spearman's correlation is easier to calculate than Kendall's tau.

6 Results, Evaluation and Discussion

The current research on blogosphere depends on statistics obtained from multiple sources. These statistics are further used in analysing the individual as well as a group of bloggers to identify various types of influential bloggers and study their characteristics. As the concept of influence is a subjective phenomenon, thus, the research domain of influential users finding lacks benchmark datasets and ground truth [42]. This lack of ground truth makes it difficult to verify and validate the proposed model [43]. Therefore, we are building a logical mechanism to analyse the proposed model from different perspectives to bring more clarity into the research. The work is the closely related to and in continuation of the already published research [25, 36]. This section is divided into the following sub-sections:

1. Evaluation on the basis of each single feature
2. Comparative analysis with the baseline (BP & BI Index)
3. Evaluation on the basis of the Modules
4. Comparative analysis between MIBSA ranking and Expected ranks of top bloggers

6.1. Evaluation on the basis of each (Single) feature

Table 3 lists the top-10 influential bloggers on the basis of each single feature (of activity and recognition modules). First three features; number of blog posts, average length of blog posts and sentiment analysis of the blog posts are taken from the activity module. Similarly, the features; number of inlinks, number of comments, number of distinct commenters and number of outlinks are taken from the recognition module. Overall influential users are the bloggers showing top ranks in most of the features of activity and recognition modules. We shall compare the top overall influential bloggers from table 3 with the top bloggers identified by the MIBSA model. The comparison is another way of understanding the behaviour of the model to verify that the proposed model is working in an expected manner. Murph D., is the top blogger with respect to number of blog posts, sentiment score of blog posts, number of the inlinks and number of comments. Murph D., is showing higher productivity and recognition in the social community. No other blogger enjoys such higher rankings in the entire table. Therefore, according to the overall feature based results analysis in table 3, Murph D., is the top influential user. Block R., is ranked 3rd with respect to the number of blog posts, 2nd with respect to sentiment score of blog posts, 6th with respect to the number of the inlinks and 3rd with respect to the number of comments and 8th with respect to number of outlinks (higher originality) respectively. However, with respect to the number of unique commenters, Block R., gets the 8th position in the top-10 list. Therefore, overall feature based results analysis ranks Block R. as the 5th influential user.

Table 3 Top influential bloggers with respect to each single feature

	Features of Activity Module			**Features of Recognition Module**				**Overall Influential users**
	Blog Posts	**Avg. Post length (Eloquence)**	**Sentiment score**	**Inlinks**	**Comments**	**Unique commenters**	**Outlinks**	
1	*Murph D.*	Stern J.	*Murph D.*	*Murph D.*	*Murph D.*	June L.	*Murph D.*	*Murph D.*
2	Rojas P.	Lai R.	June L.	June L.	June L.	*Topolsky J.*	Miller R.	June L.
3	*Block R.*	Miller R.	*Block R.*	*Miller P.*	*Block R.*	Lai R.	Flatley J.	*Miller P.*
4	*Miller P.*	*Topolsky J.*	*Miller P.*	*Ricker T.*	*Topolsky J.*	*Murph D.*	Blass E.	*Topolsky J.*
5	Melanson D.	Flatley J.	Rojas P.	*Patel N.*	*Miller P.*	Savov V.	*Patel N.*	*Block R.*
6	*Ricker T.*	June L.	Melanson D.	*Block R.*	*Patel N.*	*Ziegler C.*	Rojas P.	June L.
7	*Patel N.*	*Ziegler C.*	*Ricker T.*	*Topolsky J.*	*Ricker T.*	*Miller P.*	*Topolsky J.*	*Ricker T.*
8	June L.	Stevens T.	*Patel N.*	Melanson D.	Melanson D.	*Block R.*	*Block R.*	*Patel N.*
9	*Topolsky J.*	Savov V.	*Topolsky J.*	*Ziegler C.*	*Ziegler C.*	*Patel N.*	*Ricker T.*	*Ziegler C.*
10	*Ziegler C.*	Blass E.	*Ziegler C.*	Miller R.	Rojas P.	*Ricker T.*	June L.	Rojas P.

On the other hand, June L., shows an unusual behaviour. With smaller numbers of blog posts, June L., is ranked 2nd with respect to the inlinks, comments and sentiment scores of blog posts respectively, 6th with respect to eloquence in the blog posts, and top with respect to the number of unique commenters. June L., also show higher originality (ranked tenth with respect to number of outlinks). With fewer numbers of blog posts, June L., show higher recognition than any other blogger and higher levels of productivity. Therefore, overall feature based results analysis ranks June L. as the second most influential

user. The comparison of Miller P., and Ricker T., is interesting being ranked 4th and 6th with respect to the number of blog posts, 4th and 7th with respect to sentiment score of blog posts, 3rd and 4th with respect to the number of inlinks, 5th and 7th with respect to the number of comments, 7th and 10th with respect to the number of unique commenters. Also, Miller P., shows higher originality than Ricker T. Therefore, according to overall feature based results analysis, Miller P., is expected to get a higher rank (3rd) than Ricker T. (7th). On the other hand, Patel N. and Topolsky J., are ranked 7th and 9th with respect to the number of blog posts, 8th and 9th with respect to sentiment score of blog posts, 5th and 7th with respect to the number of inlinks. However, Topolsky J., is ranked higher than Patel N., with respect to the following features; eloquence in the blog posts (3rd and 15th), number of comments (4th and 6th) and number of unique commenters (2nd and 9th) and number of outlinks (7th and 5th). With smaller number of blog posts, Topolsky J., shows higher recognition and eloquence than Patel N. and almost equal sentiment score for the blog posts. Therefore, according to the overall feature based results analysis Topolsky J. (fourth) is expected to get a higher rank than Patel N (eighth).

Table 4 Top influential bloggers with respect to each single feature

	BI-Index	**BP-Index**	**Activity Module**	**Recognition Module**	**MIBSA**
1	*Murph D.*	*Murph D.*	*Murph D.*	June L.	*Murph D.*
2	*Ziegler C.*	*Ziegler C.*	Block R.	*Murph D.*	Miller P.
3	*Savov V.*	*Savov V.*	Miller P.	Ricker T.	Ricker T.
4	Miller P.	Ricker T.	Rojas P.	Miller P.	Block R.
5	Flatley J.	Topolsky J.	*Melanson D.*	Patel N.	June L.
6	Stevens T.	Miller P.	Ricker T.	Topolsky J.	*Melanson D.*
7	Stern J.	Stevens T.	*Ziegler C.*	Block R.	Patel N.
8	Ricker T.	Flatley J.	Patel N.	*Melanson D.*	Topolsky J.
9	Topolsky J.	Miller R.	Topolsky J.	*Ziegler C.*	*Ziegler C.*
10	Miller R.	Stern J.	Blass E.	Miller R.	Rojas P.

6.2. Comparison with baseline (BP & BI Index)

In this section, we compare the proposed model (MIBSA) with the baseline. The comparison is in two folds, first, we compare the baseline metrics (BP & BI index) with each module (i.e. activity and recognition) and second, we compare the baseline metrics with the proposed model to prove that MIBSA shows higher accuracy in ranking the top influential bloggers. Also, we try to investigate the reason of ranking differences between MIBSA model and the baseline.

Table 4 shows the top bloggers with respect to baseline metrics, module of activity and recognition and the MIBSA model respectively. Darren M., is ranked as the top blogger baselines (BP & BI index) and MIBSA model as Darren M., show higher recognition and productivity than any other blogger. Therefore, the baseline and MIBSA ranks Darren M., as the top blogger Ziegler C., is ranked 2nd by the baseline, whereas Ziegler C., is ranked 7th by the activity module and 9th by the recognition module. The MIBSA model ranks Ziegler, C., as the 9th influential blogger in the top bloggers list as MIBSA is based on the weighted sum of the modules where weights represent the importance of each module in determining the top influential bloggers. In feature wise discussion in section 6.1, we anticipated that

Ziegler, C., get a lower rank. So, this result is in accordance with the feature wise analysis performed in table 3 and clearly shows the flaws in the baseline metrics.

Table 5 A comparison between MIBSA and baseline metrics

	OSim	**Spearman's Correlation**	**Kendall's Correlation**
BP Index vs. MIBSA	0.5	0.55	0.37
BI Index vs. MIBSA	0.5	0.2	0.2

The baseline metrics, BP and BI index, rank Savov V. as the 3rd influential blogger, whereas Savov V. is not even among the top-10 list of modules (activity and recognition) and the MIBSA model. According the single feature analysis in table 3, Savov V., is ranked 13th with respect to the features Inlinks, and 16th with respect to the feature number of comments. This shows that baseline metrics give much too importance to the number of inlinks and comments. On the other hand, MIBSA includes five other features from blogosphere besides inlinks and comments to compute a blogger's productivity and recognition in the social community with higher accuracy. Similarly, Melanson D., is ranked 5th, 8th and 6th according to the modules and the MIBSA model respectively. Melanson D., is placed in similar positions by a number of single features in table 3, therefore, it is evident that MIBSA model gives clearer, realistic and accurate blogger rankings than the baseline metrics. On the other hand, June L., is ranked as the 5th influential blogger by the MIBSA model. With fewer blog posts, June L. shows higher recognition (higher number of inlinks, comments and unique commenters) within the social community, however, June L., is not among the top bloggers list of baseline metrics. This shows that the baseline

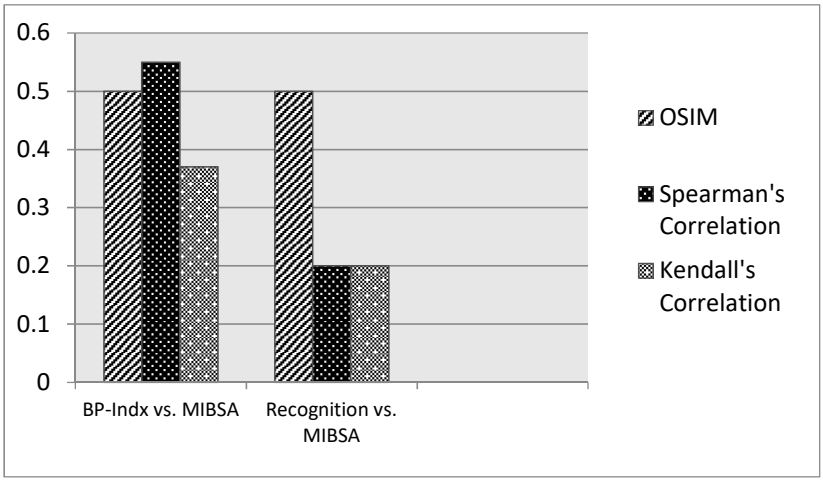

Figure 2 A comparison between MIBSA and baseline metrics

does not take into account the number of users inspired to write comments on the blog posts of influential bloggers.

From table 5, high values of the performance evaluation measure OSim show that the overall results of MIBSA model and the baseline metrics are similar which means the results are valid. The comparison between MIBSA and the BP index results in five common bloggers as shown in table 4. Similarly, the

comparison between MIBSA and BI index has also resulted in five common bloggers. The performance evaluation measure OSim in table 5 validates our comparative analysis. However, Spearman and Kendall correlation values are very small which indicates the different ranking orders provided by MIBSA. This is because the MIBSA is based on a modular approach and uses seven features (more features than any of the existing models and the baseline metrics) from the blogosphere. Figure 2 graphically represents the comparative analysis given in table 5.

6.3. Evaluation using the Modules of Activity and Recognition)

As mentioned earlier that the activity module measures blogger productivity, whereas the recognition module measures the blogger reputation and influence in the social community. The MIBSA model integrates the activity and recognition modules to identify the productive and influential bloggers. Therefore, each module plays a significant role in identifying the top influential bloggers. Unlike the baseline metrics which uses only inlinks and comments, the proposed model is based on a comprehensive set of features from the blogosphere. Table 6 presents a comparison between the MIBSA and its modules to prove the importance of each module in finding influential bloggers. The comparative analysis in table 6 shows that the overall rankings of activity and recognition module are consistent but there are few exceptions. For instance, June L. is ranked as the top blogger by the recognition module, whereas June L. is not among the list of top bloggers of the activity module. On the other hand, MIBSA model has ranked June L. among the top five influential bloggers. This is because MIBSA gives importance to both blogger productivity and recognition while ranking the top bloggers, whereas the individual modules either take into account blogger productivity or blogger recognition and influence in the social community.

Table 6 Comparison between complete model with individual modules

Activity	**Recognition**	**MIBSA**
Murph D.	*June L.*	Murph D.
Block R.	Murph D.	Miller P.
Miller P.	Ricker T.	Ricker T.
Rojas P.	Miller P.	*Block R.*
Melanson D.	Patel N.	*June L.*
Ricker T.	Topolsky J.	Melanson D.
Ziegler C.	*Block R.*	Patel N.
Patel N.	Melanson D.	Topolsky J.
Topolsky J.	Ziegler C.	Ziegler C.
Blass E.	Miller R.	*Rojas P.*

For smaller activity in the social community, June L. has a very high number of inlinks and influences many individuals (unique commenters) to write comments. This is one of the reasons of integrating each module into the MIBSA model to show higher accuracy. Similarly, Rojas P., and Blass E., are fourth and tenth productive bloggers. Despite higher productivity, Rojas P. and Blass E., show smaller influence in the social community. Therefore, Rojas P. (seventeenth) and Blass E., (twenty-one) are not among the list of top recognized bloggers. However, the MIBSA model ranks Rojas P., as the tenth influential blogger as Rojas P., show very high productivity, but moderate influence and Blass E., is not in the list of top bloggers as Blass E., show high productivity, but smaller influence in the social

community. On the other hand, Block R., the second productive and seventh recognized blogger respectively. So, Block R., gets an overall fourth position in the ranking list of MIBSA model. This proves the point that, unlike individual modules, MIBSA model shows higher accuracy in identifying the active as well as the influential bloggers in the social community.

Table 7 Comparative analysis between complete model with individual modules

	Spearman C orrelation	**OSim**	**Kendall Correlation**
Activity vs. MIBSA	0.8	0.9	0.68
Recognition vs. MIBSA	0.34	0.9	0.37

Now, we present a comparative analysis between MIBSA model and both the modules using the performance evaluation measures of OSim, Kendall correlation and Spearman's Correlation. It can be seen in table 7 that MIBSA shows higher correlation (0.8), higher similarity (0.9) and higher variance (0.68) with the activity module. MIBSA also shows higher similarity (0.9), but lower correlation (0.34) and lower variance (0.37) with the recognition module.

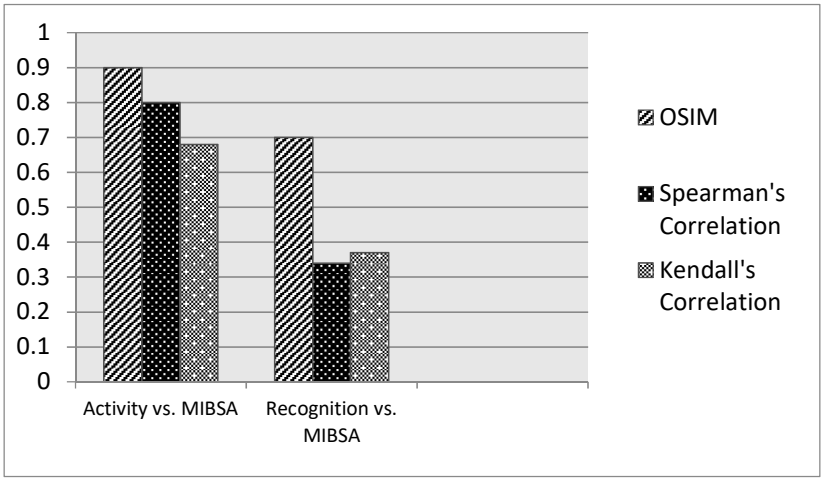

Figure 3 A comparative analysis between MIBSA and each module

Figure 3 represents the comparative analysis using performance evaluation measures. MIBSA shows higher correlation, higher similarity and high variance with the activity module than recognition module.

6.4. Use and Effects of Weights

As the domain of influential bloggers lacks the ground truth, therefore, we establish a logical mechanism to validate the proposed model. One method of such validation is the comparative analysis given in Table 8 in which combined effect of modules of activity and recognition using different weights (weights being the importance of each module in determining the influential bloggers) are shown as compared to Overall Influential Users to verify that MIBSA ranking is rational and consistent. In table 8, ω_{act} and ω_{rec} are the weights of activity and the recognition modules respectively. It is notable in Table 9 that

optimal correlation values are achieved, where the MIBSA ranking best match the expected influential users when ω_{act}=0.4 and ω_{rec}= 0.6.

Table 8 Comparison between MIBSA and the expected influential bloggers by varying weights

	Expected top-bloggers	**ω_{act}=0.1 ω_{rec}=0.9**	**ω_{act}=0.2 ω_{rec}=0.8**	**ω_{act}=0.3 ω_{rec}=0.7**	**ω_{act}=0.4 ω_{rec}=0.6**	**ω_{act}=0.5 ω_{rec}=0.5**	**ω_{act}=0.6 ω_{rec}=0.4**	**ω_{act}=0.7 ω_{rec}=0.3**	**ω_{act}=0.8 ω_{rec}=0.2**	**ω_{act}=0.9 ω_{rec}=0.1**
1	Murph D.	June L.	June L.	June L.	Murph D.	Miller P.	Murph D.	Murph D.	Murph D.	Murph D.
2	June L.	Murph D.	Murph D.	Murph D.	Miller P.	Murph D.	Miller P.	Miller P.	Miller P.	Miller P.
3	Miller P.	Ricker T.	Ricker T.	Ricker T.	June L.	Block R	Block R	Block R	Block R	Block R
4	Topolsky J.	Miller P.	Miller P.	Miller P.	Block R.	June L.	Melanson D.	Rojas P.	Rojas P.	Rojas P.
5	Block R.	Patel N.	Patel N.	Patel N.	Ricker T.	Melanson D.	June L.	Melanson D.	Melanson D.	Melanson D.
6	June L.	Topolsky J.	Topolsky J.	Topolsky J.	Patel N.	Ricker T.	Ricker T.	Ricker T.	Ricker T.	Ziegler C.
7	Ricker T.	Block R.	Block R.	Block R.	Melanson D.	Patel N.	Rojas P.	Ziegler C.	Ziegler C.	Ricker T.
8	Patel N.	Melanson D.	Melanson D.	Melanson D.	Topolsky J.	Ziegler C.	Patel N.	June L.	June L.	June L.
9	Ziegler C.	Ziegler C.	Ziegler C.	Ziegler C.	Ziegler C.	Topolsky J.	Ziegler C.	Patel N.	Patel N.	Topolsky J.
10	Rojas P.	Miller R.	Miller R.	Miller R.	Rojas P.	Rojas P.	Topolsky J.	Topolsky J.	Topolsky J.	Blass E.

Table 9 Correlation analysis between the expected influential bloggers and MIBSA (by varying weights)

	ω_{act}=0.1 ω_{rec}=0.9	**ω_{act}=0.2 $\omega_{rec}pec$ =0.8**	**ω_{act}=0.3 ω_{rec}=0.7**	**ω_{act}=0.4 ω_{rec}=0.6**	**ω_{act}=0.5 ω_{rec}=0.5**	**ω_{act}=0.6 ω_{rec}=0.4**	**ω_{act}=0.7 ω_{rec}=0.3**	**ω_{act}=0.8 ω_{rec}=0.2**	**ω_{act}=0.9 ω_{rec}=0.1**
Expected Influential bloggers	0.58	0.63	0.69	0.81	0.75	0.71	0.65	0.64	0.6

7 Conclusion and Future Research Directions

In this paper, we address the problem of finding influential bloggers from an online social community. We propose a modular framework based on seven features. One of the key features of the proposed framework is that it uses the modules of activity and recognition to identify the blogger not only productive, but also influential in their blogging communities. The results prove that the modules play a significant role in identifying the top influential bloggers with higher accuracy than the baseline metrics and the novel feature, Sentiment analysis of the blog posts, effectively captures the blogger productivity by taking into account the (non-sentimental) objective information a blogger writes in the blog posts.

Therefore, sentiment analysis should be considered an important feature in identifying the influential bloggers.

A main challenge in the relevant literature is the absence of benchmark datasets and lack of standards [42] which makes it difficult to verify and validate the proposed models. Due to the dynamic nature of blogging platforms, it is difficult to keep track of temporal aspects a blogger's influence. For future work, we aim to extend the proposed model by including temporal based features and investigate the impact of temporal features on the overall accuracy in identifying the influential bloggers. Also, we plan to quantify the blogger influence respect to various topics.

References

1. Munger, T. and Zhao, J., Identifying influential users in on-line support forums using topical expertise and social network analysis. in Proceedings of IEEE/ACM ASONAM Conference, (Paris, 2015).
2. Gliwa, B., Koźlak, J., Zygmunt, A. and Demazeau, Y., Combining Agent-Based and Social Network Analysis Approaches to Recognition of Role Influence in Social Media.in Proceedings of 14th PAAMS Conference, (Sevilla, 2016).
3. Bouguessa, M. and Romdhane, L. B. Identifying Authorities in Online Communities. ACM Transactions on Intelligent Systems and Technology, 6(3). 2015. 1-23.
4. Gao, K., Xu, H. and Wang, J. Mining blogs and forums to understand the use of social media in customer co-creation. Computer Journal. 58(9), 2015, 1909-1920.
5. Kao, L.-J. and Huang, Y.-P., Mining Influential Users in Social Network. in Proceedings of IEEE SMC Conference, (Hong Kong, 2015).
6. Singer, Y., How to Win Friends and Influence People, truthfully: influence maximization mechanisms for social networks. in Proceedings of ACM WSDM Conference, (Seattle, 2012)
7. Agarwal, N., Liu, H., Tang, L. and Yu, P. S. Modelling blogger influence in a community. Social Network Analysis and Mining, 2(2), 2012, 139-162.
8. Tarokh, M. J., Arian, H. S. and Speily, O. R. B. Discovering Influential Users in Social Media to Enhance Effective Advertisement. Advances in Computer Science: An International Journal (ACSIJ), 4(5), 2015, 23-28.
9. Bilanakos, C., Sotiropoulos, D. N., Georgoula, I. and Giaglis, G. M., Optimal Influence Strategies in Social Networks. in Proceedings of the IEEE/ACM ASONAM Conference, (Paris, 2015).
10. Akritidis, L., Katsaros, D. and Bozanis, P. Identifying the Productive and Influential Bloggers in a Community. IEEE Transaction on System, Man, Cybernetics, Part C, 41(5), pp. 759-764, 2011.
11. Haveliwala, T., Topic-sensitive PageRank. in Proceedings of WWW Conference, (Honolulu, 2002).
12. Agarwal, N., Liu, H., Tang, L. and Yu, P. S., Identifying the Influential Bloggers in a Community. in Proceeding of WSDM Conference, (Stanford, 2008).
13. Brin, S. and Page, L., The anatomy of a large-scale hyper-textual Web search engine. in Proceeding of 7th international conference on World Wide Web, (Brisbane, 1998).
14. Qureshi, M. A., Younus, A., Saeed, M. and Touheed, N., Identifying and ranking topic clusters in the blogosphere. in Proceeding of ACM COLING Conference, (Beijing, 2010).
15. Agarwal, N., A study of communities and influence in blogosphere. in Proceeding of 2nd SIGMOD IDAR Workshop, (Vancouver, 2008).
16. Akritidis, L., Katsaros, D. and Bozanis, P., Identifying Influential Bloggers: Time Does Matter. in Proceeding of WI-IAT Conference, (Milan, 2009).

17. Chen, T., Xu, R., He, Y., Xia, Y., and Wang, X. Learning User and Product Distributed Representations Using a Sequence Model for Sentiment Analysis. IEEE Computational Intelligence Magazine, 11(3), pp. 34-44, 2016.
18. Cambria, E., Poria, S., Bajpai, R. and Schuller, B., SenticNet 4: A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives. in Proceeding of 26th International Conference on Computational Linguistics, pp. 2666-2677, (Osaka, 2016).
19. Cambria, E. Affective computing and sentiment analysis," IEEE Intelligent Systems. 31(2), pp. 102-107, 2016.
20. Zadeh, A., Zellers, R., Pincus, E. and Morency, L.-P. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. IEEE Intelligent Systems, 31(6), pp. 82-88, 2016.
21. Aziz, M. and Rafi, M., Identifying influential bloggers using blogs semantics. in Proceedings IEEE FIT Conference, (Abbottabad, 2010).
22. Moh, T.-S. and Shola, S. P., New Factors for Identifying Influential Bloggers. in Proceedings of IEEE International Conference on Big Data, (Silicon Valley, 2013).
23. Gliwa, B. and Zygmunt, A. Finding Influential Bloggers. International Journal of Machine Learning and Computing, 5(2), 2015, 127-131.
24. Alghobiri, M., Ishfaq, U., Khan, H. U. and Malik, T. A., Exploring the role of sentiments in identification of active and influential bloggers. in Proceedings of International Conference on Computer Science and Communication Engineering, (Durres, 2015).
25. Ishfaq, U., Khan, H. U. and Iqbal, K., Modelling to find the top Bloggers using Sentiment Features. in Proceeding of IEEE ICE CUBE 2016, (Quetta, 2016).
26. Khan, H. U., Daud, A., and Malik, T. A. MIIB: A Metric to Identify Top Influential Bloggers in a Community. PLoS ONE, 10(9), pp. 1-15, 2015.
27. Li, Y.-M., Lai, C.-Y. and Chen, C.-W., Identifying bloggers with marketing influence in the blogosphere. in Proceeding of ICEC Conference, (Taipei, 2009).
28. Li, Y. M., Lai, C. Y. and Chen, C.-W. Discovering influencers for marketing in the blogosphere. Information Sciences, 181(23), pp. 5143-5157, 2011.
29. Goyal, A., Bonchi, F. and Lakshmanan, L. V.S., Discovering leaders from community actions. in Proceeding of CIKM Conference, (Nepa Valley, 2008).
30. Kayes, I., Qian, X., Skvoretz, J. and Iamnitchi, A., How Influential Are You: Detecting Influential Bloggers in a Blogging Community. in Proceeding of SocInfo Conference, (Lausanne, 2012).
31. Bui, D.-L., Nguyen, T.-T. and Ha, Q.-T., Measuring the Influence of Bloggers in their Community Based on the H-index Family. in 2nd International Conference on Computer Science, Applied Mathematics and Applications, (Budapest, 2014).
32. Shang, J., Zhou, S., Li, X., Liu, L., and Wu, H. CoFIMCoFIM: A community-based framework for influence maximization on large-scale networks. Knowledge-Based Systems, 117, pp. 88-100, 2017.
33. Gong, M., Song, C., Duan, C., Ma, L. and Shen, B. An Efficient Memetic Algorithm for Influence Maximization in Social Networks. IEEE Computational Intelligence Magazine, 11(3), pp. 22-33, 2016.
34. Sumith, N., Annappa, B. and Bhattacharya, S. Social network pruning for building optimal social network: A user perspective. Knowledge-Based Systems, 117, pp. 101-110, 2017.
35. Calderoni, F., Brunetto, D. and Piccardi, C. Communities in criminal networks: A case study. Social Networks, 48, pp. 116-125, 2017.
36. Khan, H. U., Daud, A. Finding the top influential bloggers based on productivity and popularity features. New Review of Hypermedia and Multimedia, doi: 10.1080/13614568.2016.1236151, 2016.

37. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. and Kappas, A. Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), pp. 2544–2558, 2010.
38. Vural, G., Cambazoglu, B. B., and Senkul, P., Sentiment-focused web crawling. in Proceedings of the 21st ACM International Conference on Information and Knowledge Management, (Maui, 2012).
39. Giannopoulos, G., Weber, I., Jaimes, A. and Sellis, T., Diversifying User Comments on News Articles. in Proceedings of the WISE Conference; (Paphos, 2012).
40. Thelwall, M., Buckley, K., Paltoglou, G. Sentiment strength detection for the social Web. Journal of the American Society for Information Science and Technology, 63(1), 2012, 163-173.
41. Glass, A. L. Cognition A Neuroscience Approach. Cambridge University Press, 2016.
42. Agarwal, N., Mahata, D. and Liu, H. Time and Event Driven Modeling of Blogger Influence, Encyclopaedia of Social Network Analysis and Mining (ESNAM). Alhajj, Reda; Rokne, Jon (Eds.). Springer, 2014.
43. Khan, H. U., Daud, A., Ishfaq, U., Amjad, T., Aljohani, N., Abbasi, R. A. and Alowibdi, J. S. Modelling to identify influential bloggers in the blogosphere: A survey. Computers in Human Behavior. 68, pp. 64-82, 2017.