International Journal on Artificial Intelligence Tools
© World Scientific Publishing Company

Fuzzy Information Diffusion in Twitter by Considering User's Influence

Andreas Kanavos
Department of Computer Engineering and Informatics, University of Patras,
GR 265-00, Greece.
kanavos@ceid.upatras.gr

Ioannis E. Livieris
Department of Mathematics, University of Patras,
GR 265-00, Greece.
livieris@gmail.com

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Does a post with specific emotional content that is posted on Twitter by an influential user have the capability to affect and even alter the opinions of those who read it? Accordingly, "influential" users affected by this post can then affect their followers so that eventually a large number of users may change their opinions about the subject the aforementioned post was made on? Social Influence can be described as the power or even the ability of a person to yet influence the thoughts and actions of other users. So, User Influence stands as a value that depends on the interest of the followers (via replies, mentions, retweets, favorites). Our study focuses on identifying such phenomena on the Twitter graph of posts and on determining which users' posts can trigger them. Furthermore, we analyze the Influence Metrics of all users taking part in specific discussions and verify the differences among them. Finally the percentage of Graph cover when the diffusion starts from the "influential" users, is measured and corresponding results are extracted. Hence, results show that the proposed implementations and methodology can assist in identifying "influential" users, that play a dominant role in information diffusion.

Keywords: Graph Mining, Fuzzy Sets, Information Diffusion, Knowledge Extraction, Social Media Analytics, User Influence, Sentiment Analysis

1. Introduction

Given the global spread of social media usage over the last years, it has become clear that microblogging platforms, like Twitter, are being used as a means to either express or be informed about the personal views of many users on various subjects. Additionally, the simplification of posting and sharing content on these platforms has led to the creation of vast amounts of information on a daily basis, so that the following question arise: "*Is it possible to analyze all this content in such a way that certain of its aspects, like sentiment or informational values, can be obtained in the* *form of meta-data? Can this meta-data be processed and used in such a way that future information from the exact source will be affected in a desired way?".*

Tweets have the form of small sentences up to 140 characters long, that can be accompanied by a photo. Many Twitter users post a vast amount of information to the public or selected circles of their contacts on a daily basis and a logical assumption, that specific analysis of this information can thus lead to results that convey knowledge about certain aspects of the users' characters, such as their beliefs, emotional states or behavior patterns could be established. The users' posts in Twitter, unlike in other networks, have some special characteristics. The short length that the posts are restricted to have, results in more expressive emotional statements. This enormously continuous stream of Twitter data posts, reflects users' opinions and reactions to phenomena from political events all over the world to consumer products [31].

Analyzing tweets and recognizing their emotional content is a very interesting and challenging topic in the microblogging area. It is necessary for deeper understanding of people's behavior, but simultaneously for describing public attitude towards different events and topics as well. Therefore it could be helpful in predicting the spread of posts and information diffusion in the network. Such an analysis can have a great impact on platforms like Twitter, where users' posts hold more emotional content than other platforms due to the existing restriction in the length of users' posts.

Lately, great interest for the results of research in this field has been shown not only by members of academic communities all across the world, but also by companies that care about users' opinions and reactions to various phenomena from political events to consumer products. Hence, it is of paramount significance to detect the specific Twitter users, whose opinions are expressed through their tweets and have the most influence on the opinions, beliefs and sentiments of the rest of the users. Such users are often called "influential" and play a key part in actively shaping the general mood around certain subjects discussed on Twitter. Just like in real life communities, most social media users tend to agree with those among them who have high social status or specific traits, like extensive knowledge on some field or fluency of speech.

PageRank has also served as a model for computing digital influence. For instance, TwitterRank [34] factors in the similarity between accounts as well as edge patterns. The authors claim in the same work that TwitterRank outperforms other related algorithms such as In-degree, PageRank, and Topic-sensitive PageRank. Another algorithm derived from PageRank is TunkRank[a], which recursively computes the digital influence of Twitter account.

The primary contribution of this work is to highlight the propagation of influential users' posts on a Twitter graph and thus, to effectively predict the alternation of the emotional bias influence of future posts by other users on certain subjects.

[a] `https://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank`

Our proposed method analyzes the sentiment of all the posts on a given subject of discussion on Twitter and calculates the positive or negative bias of that particular discussion. Additionally, we propose the evaluation of all the users who took part in the discussion, based on their influence metrics, as well as the forming of a group having the most influential users. Another significant contribution constitutes the Fuzzy Random Graph, where each edge between two Twitter users shares an equal probability of contact, thus being independent from the other pairs. Finally, the Rate of Covered Users (or percentage of Graph cover) in case the diffusion starts from the "influential" users, is presented. Hence, we can state that our assumption of the "influential" users playing a dominant role in information diffusion, has been verified. All the aforementioned features utilize a dynamically refined framework that takes into consideration features of the Twitter network as well as fuzzy methodologies, e.g. random graph.

The remainder of this work is structured as follows: Section 2 presents background topics in influence estimation, complex networks and fuzzy sets. Section 3 focuses on our proposed methodology. Section 4 details the implementation of the system, while Section 5 presents the evaluation of the experiments and the results gathered from our proposed method. Finally, Section 6 recapitulates the conclusions and enumerates directions for future work.

2. Related Work

2.1. *User Influence*

Commercial companies and associations could exploit Twitter for marketing purposes, as it provides an effective medium for propagating recommendations through users with similar interests. Moreover, viral marketers could exploit models of user interaction to spread their content or promotions quickly and widely [20].

Analyzing and determining the opinion of the public regarding news and topics in social media, constitutes a rather active research area with an abundance of works and a wide range of applications varying from government agencies to financial and commerce companies [18,30]. For example, governments would like to have an estimation of the public's attitude towards their political actions, while commercial companies would like to know how users comment on products and spot specific target groups [33]. Twitter and other social media platforms have actually facilitated citizens' protests against government and ultimately even revolutions. Therefore, governments and law enforcement agencies are now placing a greater emphasis on examining and controlling the structure of social networks [15]. Authors study the emotion synchronization for a group of individuals in a social network; they try to model human emotion as a fractional order system and in following, a decentralized potential field-based function is developed, so as to ensure that the emotion states of all individuals asymptotically converge to a common equilibrium while maintaining social bonds. They have found that social network users tend to put greater trust in a close friend than some random person, and thus, can be more easily influenced by this close friend.

In the work presented in [5], authors investigate how ratings on a piece of content affect its author's future behavior. Notably, authors studied whether community feedback regulates the quality and quantity of a user's future contributions in a way which benefits the community. Moreover, they found that negative feedback leads to significant changes in the author's behavior, which are much more salient than the effects of positive feedback.

The significance of estimating the behavior of a user in a Twitter network with respect to the percentage of their "influence" has already been discussed and several modifications have been proposed [12,14,16,17]. Specifically, in [14], the authors extend the notion of influence from users to networks and considered personality as a key characteristic for identifying influential networks. Their proposed system creates influential communities in a Twitter network graph by considering users' personality; the authors used an existing modularity-based community-detection algorithm and later extend it by inserting a pre-processing step which eliminates graph edges based on users' personality. In [16,17], a methodology for estimating the importance and the influence of a user in a Twitter Network is described. Moreover, the authors propose a schema where users are represented by nodes and the edges, which connect these nodes, represent the relations of Follower to Following introduced by Twitter. The metric type for the influence is determined with the use of well known features of Twitter, including the Frequency of users posts, the number of Followers, etc.

Furthermore, the analysis presented in [27] is also worth mentioning. The creation of *Influence Tracker*, a publicly available website where anyone can rate and compare the recent activity of any Twitter account, is presented. This Influence Metric depends on the number of Tweets, Followers, Following as well as Frequency. The main difference of our proposed methodology, regarding influence estimation, is the emotional analysis which we take into consideration along with the user information, e.g. Twitter analytics.

2.2. *Complex Networks*

The epidemic spreading achieves high attention towards the understanding of the unfolding of dynamical processes in complex networks. Authors in [24] reviewed and presented various solved and open problems in the development, analysis and finally in the control of epidemic models. A coherent and comprehensive review of the vast research activity concerning epidemic processes is presented in [25]. A similar work is presented in [21], where authors study the spreading of infections in complex heterogeneous networks based on an SIRS (Susceptible - Infectious - Recovered - Susceptible) epidemic model with birth and death rates. Specifically, the SIRS model is considered in the clustered scale-free networks, in order to examine the effect of network community structure on the epidemic dynamics.

Generally, the modelling of complex networks provides a lot of information and can assist in thoroughly understanding their characteristics. It enables the simulation, analysis as well as prediction of the behavior of processes taking place in them, such as diffusion or information retrieval.

The model of Erdös and Rényi [9] is one of the first network models that represents the random graph. Notably, the random graph model is characterized by the number of nodes and by the probability of connecting two arbitrarily selected nodes. Each one of the pairs of nodes is associated with an equal probability to each other, independent of the other pairs [7,9]. The Erdös - Rényi model has been accepted due to its properties which facilitate the network modeling. Random graphs do not reflect the structure of real networks, because the degrees of random graph nodes follow the Poisson [10], instead of the power-law distribution[b] [1]. Thus, the Erdös - Rényi model does not reflect the effect of clustering, while the random graph can be considered a perfect model choice for studying complex networks [9].

2.3. *Fuzzy Sets*

Zhang et al. [36] recalled the definition of intuitionistic fuzzy sets and the principle of information diffusion. Moreover, they introduced some constraint conditions for satisfying the requirements of conversion from intuitionistic fuzzy sets into fuzzy sets and deduced a new conversion operator which seeks to obtain one and only fuzzy set for each intuitionistic set. Additionally, they provided a new fuzzy entropy formula for intuitionistic fuzzy sets, based on a distance which is similar to the consideration for ordinary fuzzy sets. Their preliminary numerical experiments demonstrated that the proposed entropy formula has a simple format and is convenient for operation.

Lin and Liao [22] studied the privacy preserving publishing of social network data and presented a new methodology to model the social network data as directed graphs with signed edge weights; formally define privacy, attack models for the anonymization problem. Their main advantage of their graph clustering algorithm is that it can effectively group similar graph nodes into clusters with minimum cluster size constraints. They presented a series of experiments to evaluate the effectiveness and utility of their approach on anonymizing social network data.

Zhang et al. [37] presented a new framework for improving the credibility and the speed of service diffusion in social network. More analytically, they proposed a new method which identifies nodes with better trust values and better diffusion characteristics by utilizing the trustworthy relationship among social network nodes. They evaluated the proposed method against a non-centralized control model such as the flooding model and the random walk model. Their reported numerical results revealed that the proposed method can efficiently identify the optimizational trustworthy nodes in less computational time.

In more recent works, Keshavarz et al. [19] considered a novel approach for generating sentiment lexicons and assigning numerical scores to words present in text, based on the frequencies of words in positive and negative text. These lexicons are

[b] `http://www.necsi.edu/guide/concepts/powerlaw.html`

used to calculate meta-level features, which can be used alongside the features of other lexicons to improve their accuracy. Their experiments on six datasets illustrated the efficiency of their proposed framework.

3. Proposed Method

This section presents the architecture of the system which has been developed, in order to address the needs of the present work. Figure 1 illustrates its components, as well as the information flow between them. The main functions of this system are the social media crawler component, the tweet emotion recognition system and the influence processing method.

The proposed methodology aims to set the basis for the creation of a model for analyzing conversations on Twitter. Initially, a large number of tweets from a specific topic discussed on Twitter are gathered with the use of a social media crawler, which traverses the Twitter graph and extracts the posts made, as well as the corresponding users who addressed them, along with other information about them (number of followers, retweets, etc).

3.1. *Fuzzy Logic and Risk Estimators*

The concept of risk plays a central role in modern epidemiology. More to the point, in recent works, individuals at risk are supposedly exposed or non-exposed to a certain cause with the aim of being in following categorized as diseased or non-diseased [23,29].

In this work, a different approach is introduced, as we incorporate the concepts of fuzzy logic and individual outcomes. Individuals are assumed to be exposed to some risk factor according to certain fuzzy set membership functions and in following their response is categorized according to other fuzzy set membership functions. As a next step, risk analysis is utilized by applying fuzzy set theory as well as maximum likelihood. By introducing these aspects, fuzzy relative risk ratio under individual heterogeneity is calculated, thus giving us more realistic estimators than their classical counterparts.

One classic kind of risk estimator in epidemics is the Risk Ratio (RR) which is defined as the ratio of the conditional probability of developing a disease given that one is exposed to a certain cause $p(D \,|\, E)$, to the conditional probability of developing the disease given that one is not exposed to the cause $p(D \,|\, \overline{E})$, which is:

$$RR = \frac{p(D \,|\, E)}{p(D \,|\, \overline{E})} \tag{1}$$

Nevertheless, in case the risk ratio is defined in terms of conditional probabilities, then it is reasonable that under fuzzy logic setting it should be defined in terms of conditional possibilities. A possibility distribution function r associated with a fuzzy subset A is numerically equal to its grade of membership function μ_A [35], which is:

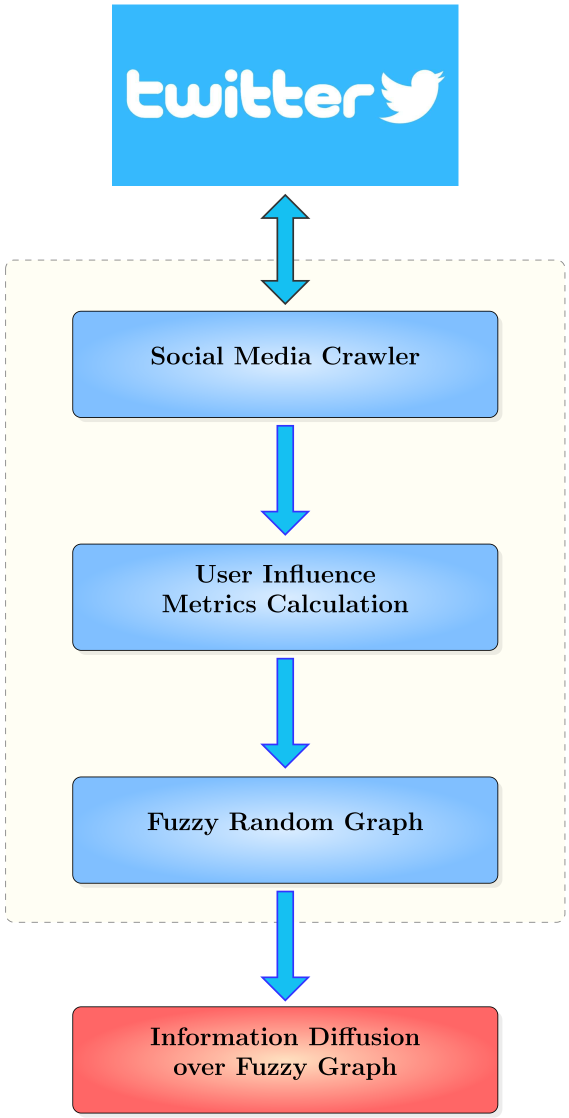

Fig. 1. Overall architecture of the proposed methodology

$$r(x) = \mu_A(x); \qquad \forall x \in X, \tag{2}$$

where X is a set serving as the universe of discourse and measure π is given by:

$$\pi(A) = \max_{x \in A} r(x). \tag{3}$$

Furthermore, another kind of risk estimator that can be introduced is the Fuzzy Risk Ratio (FRR). This ratio is defined as the ratio of the conditional possibility of developing a given disease severity d given that one is exposed to a certain level of a causal factor e, to the conditional possibility of developing disease severity d given that one is not exposed to the causal factor $\overline{e}$.

The fuzzy risk ratio estimator should then be defined in terms of conditional possibilities and thus it is expected to be proportional ($\propto$) to the ratio between the conditional possibility of developing a certain disease given that one is exposed to a suspected factor $Poss(D \mid E)$, to the conditional possibility of developing the disease given that one is not exposed to that factor $Poss(D \mid \overline{E})$, which is:

$$FRR \propto \frac{Poss(D \mid E)}{Poss(D \mid \overline{E})} \tag{4}$$

3.2. *Social Media Crawler*

In this subsection, the methodology for estimating the influence of a Twitter user in a specific Twitter graph is introduced as in [14]. The social media crawler creates a social media graph where the nodes represent the users and the edges represent the "Following" relationships among these users. In our paper, we utilize a topic-based sampling approach where tweets are collected via a number of different keyword search queries.

More specifically, the Twitter graph is generated in the following way. Initially, for a concrete #hashtag, the users and their corresponding followers, which have posted a tweet within a given time period, are retrieved. Subsequently, users that follow each other or have a common follower, are connected in order for the graph to be utilized. The process for generating the Social Media Graph is presented in a more analytical and detailed way in Algorithm 1.

3.3. *User Influence Metrics*

This subsection describes the set of metrics taken into consideration for the present manuscript. Some of them have been already proposed in our previous works [8,12,14,17]. User features are extracted by calculating and combining different measures, as proposed in the following paragraphs.

A social media crawler traverses the utilized Twitter graph of posts on a subject specified by a search query and returns a number of posts big enough to ensure opinion and emotion diversity. The time interval and the number of Tweets per keyword, as will be presented in following subsections, were two months and about 20.000, respectively. We consider zero time according to the timestamp of the first tweet containing the requested #hashtag and ending time according to the timestamp of the last such tweet by the time of query.

Specifically, as already stated in our previous works, we take into consideration some additional features that deal with interaction among different users in Twitter,

Algorithm 1 Generation of Social Media Graph

```
Require: Query/Keyword $\#q$
Ensure: The sample Graph $Users$ and lists Followers[] and Newnodes[] are computed
identify set of tweets for given $\#q$, $T = \{t_1, t_2, \ldots, t_i\}$
$\forall$ tweet $t_i \in T$
$u_i \leftarrow$ user of tweet $t_i$
Followers[$i$] $\leftarrow$ Followers of $c\,[i]$
for all $t_i \in T$ do
    $Users \leftarrow Users \cup u_i$
end for
identify set of followers of a user $u_k$, Followers[$u_k$] $= \{f_1, f_2, \ldots, f_j\}$
for all $u_k \in$ Users do
    for all $f_j \in Followers[u_k]$ do
        if $f_j \in$ Users then
            link $f_j$ with $u_k$
        else
            for all $u_l \in$ Users and $u_l \neq u_k$ do
                if $f_j \in$ Followers[$u_l$] then
                    Newnodes $\leftarrow$ Newnodes $\cup f_j$
                    link $f_j$ with $u_k$ and link $f_j$ with $u_l$
                end if
            end for
        end if
    end for
end for
Users = Users $\cup$ Newnodes
```

i.e. the number of Retweets and Replies they address as well as the Clicks, Favorites and Mentions received. Regarding Clicks, Favorites and Mentions received, these factors indicate that for some corresponding users, their posts may have a great amount of impact compared to other users [2,3,6,28]. Concerning Clicks, three different numbers are estimated in order to compute this feature, namely, Link clicks (Clicks on a URL in the Tweet), Permalink clicks (Clicks on the Tweet) and User profile clicks (Clicks on the name, username, or profile photo of the Tweet user).

Table 1 summarizes the features, which can be computed by these rankings.

User Influence Metric 1. Conversational accounts have a high number of tweets, retweets, conversations, favorites, mentions and clicks. The conversational metric μ^c is calculated as

$$\mu_k^c \triangleq |T_k| + |R_k| + |C_k| + |V_k| + |M_k| + |Cl_k| \tag{5}$$

User Influence Metric 2. Multisystemic accounts have a high number of hashtags

Table 1. Data for the k-th account

Feature	**Meaning**	**Feature**	**Meaning**
T_k	Tweet set	R_k	Retweet set
C_k	Reply set	M_k	Mention set
Fol_k	Follower set	Fr_k	Friend set
H_k	Hashtag set	V_k	Favorite set
Cl_k	Click set	F_k	Frequency

in their tweets, retweets and conversations. These accounts are probably proficient in a broad range of topics. The multisystemic metric μ^m is calculated as

$$\mu_k^m \triangleq |H_k| \tag{6}$$

User Influence Metric 3. Energetic accounts have a high number of tweets over a given time interval. This behavioral pattern likely indicates knowledge of or strong opinion about a particular topic. The energetic metric μ^e is calculated as

$$\mu_k^e \triangleq F_k \tag{7}$$

User Influence Metric 4. Popular accounts have a high number of followers. Although Twitter popularity does not necessarily correspond to optimal diffusion, highly followed users exert limited influence since they are often read. The popularity metric μ^p is calculated as

$$\mu_k^p \triangleq |Fol_k| \tag{8}$$

User Influence Metric 5. Affiliated accounts have equal number of followers and friends. This metric depends on the Jaccard similarity coefficient [11], where the result is 0 when the two sets (followers and friends) are disjoint, 1 when the two sets are equal and otherwise the output takes values between 0 and 1. In addition, this metric constitutes a commonly used indicator regarding the similarity between two sets, e.g. two sets are more similar when their Jaccard index is closer to 1 while on the other hand, they are more dissimilar when their Jaccard index is closer to 0. The affiliation metric μ^a is calculated as

$$\mu_k^a \triangleq \frac{|Fol_k \cap Fr_k|}{|Fol_k \cup Fr_k|} \tag{9}$$

User Influence Metric 6. The atomic influential metric computes the geometric mean of many of the above metrics. The assumption we take into consideration by examining the semantic property of this metric is that in most cases when a user has influential Followers, then probably he is an "influencer" too. Assuming that the path of influence is by following a user (e.g. like an edge between two nodes in a graph), a potential action between these two users could be a Retweet, a Reply, a Mention or a Favorite. The atomic influential metric μ^i is calculated as

$$\mu_k^i \triangleq \left(|\mu_k^c|\,|\mu_k^a| \log\left(1+|Fr_k|\right) \log\left(1+|Fol_k|\right)\right)^{\frac{1}{4}} \tag{10}$$

in order to capture the total online presence of an account.

3.4. *Fuzzy Random Graph*

The proposed model is based on the Erdös - Rényi random graph, as presented in Algorithm 2, where each edge between two Twitter users shares an equal probability of contact (via "Following"), thus is independent from the other pairs. While the diffusion of information in the corresponding graph was implemented according to the transient contacts model, its spread takes a fair amount of time to evolve.

It's worth mentioning that the contacts are transient and this does not necessarily last through the whole life of the epidemic, but only for specified periods.

We have slightly alternated Erdös - Rényi random graph in the way that there is some dependence between whether or not the edges are present. This approach is taken into consideration because in many practical problems, the vertices are in fact randomly positioned in some geometric space (usually Euclidean). Furthermore, two vertices are adjacent if and only if the distance between them (in some specified norm) is less than a certain quantity. These corresponding points are usually uniformly distributed in $[0, 1]^n$.

In the present manuscript, we tried to incorporate some basic notations from the random-duster model by biasing the formula for the probability of a set of edges in order to favour certain kinds of graphs arising. In this kind of models, given a graph $G = G(V, E)$ and a set of edges $A \subseteq E$, then let $c(V, A)$ denote the number of components of the graph whose vertex set is V and corresponding edge set is A. Then, the probability that the arising edges are exactly as those in A is:

$$\frac{p^{|A|}(1-p)^{|E|-|A|}q^{c(V,A)}}{\sum_{F\subseteq E} p^{|F|}(1-p)^{|E|-|F|}q^{c(V,F)}} \tag{11}$$

Observe that when $q = 1$, the Erdös - Rényi model is recovered. In addition, if $q > 1$, graphs with many components are favoured, whereas if $q < 1$, the connected graphs are favoured. Notice that the study of this model is closely linked to percolation theory and statistical physics [4].

3.5. *Tweet Sentiment Bias Calculation*

This subsection presents an additional calculation that our proposed scheme utilizes. Notably, the initial posts are analyzed and their sentiment content is extracted and evaluated using a popular sentiment analysis tool [32], entitled SentiStrength[c], while the users are sorted by their influence metric. The main reason for choosing SentiStrength is its procedures for decoding non-standard spellings and methods for boosting the strength of words, which accounted for much of its performance.

[c] `http://sentistrength.wlv.ac.uk/`

Algorithm 2 Transient Contacts Model

```
input All possible edges are considered and in following included in the graph with probability $p$.
input Variable $d$ that determines the probability that modified edges are reciprocal.
for $i = 1$ to $N$ do
    for $j = i + 1$ to $N$ do
        Set a uniform random number $u$ between $0$ and $1$.
        if $p > u$ then
            Create a reciprocal edge between node[$i$] and node[$j$].
        else
            if $d > u$ then
                Create a directed edge from node[$i$] to node[$j$].
                Create a directed edge from node[$j$] to node[$i$].
            else
                Create a directed edge from node[$i$] to node[$j$].
                Set a uniformly randomly chosen node[$h$] from the set of all nodes excluding node[$i$] and node[$j$].
                Create a directed edge from node[$h$] to node[$i$].
            end if
        end if
    end for
end for
```

This information is used by our model for recognizing whether a user is capable of changing the sentiment bias of a given conversation on Twitter. The results of the performed sentiment analysis are combinations of words and signed integers, which represent their sentiment weight. Each post is then assigned a signed integer, that stands for its sentiment bias value, calculated by combining the data created in the previous step.

The sign of the sentiment bias of each post containing such word combinations is compared to the sign of the cumulative bias of the N previous and next posts. If these two signs are equal, then the bias of the equivocal post is considered to be equal to these, otherwise it is equal to the bias calculated in the usual way by the sentiment classifier.

In addition, the Tweet Emotional Bias Calculation process is presented in Algorithm 3. The process initially takes each post of the corresponding user as input and in following it performs sentiment analysis with the use of the aforementioned sentiment bias calculator tool. Finally, this process returns a signed integer for every post that a user has addressed.

The biases calculated are now used in conjunction with the user information provided by the social media crawler, in order to pinpoint potentially influential users among those participating in the given discussion. This is done in two consecutive steps, which are followed repeatedly for each of the users taking part in the discussion. Initially, in the first step, the sign of the cumulated bias of two N-sized windows is calculated. The first window contains the N posts that precede the current user's post, while the second window contains the N posts that follow it. The window N takes values equal to 15, 30 and 60.

Algorithm 3 Tweet Sentiment Bias Calculation

```
output Sentiment in the form of an integer
for all $term \in Tweet$ do
    perform sentiment analysis $(Tweet, term, integer)$
end for
for all $term_i \in Tweet$ do
    $integer \leftarrow \sum_{j=1}^{i} integer_j.$
end for
return $(User, Tweet, integer)$
```

In the second step, the two signs calculated previously are compared. Being equal, the current user's post is not considered influential enough to alter the discussion bias and the user is marked as unexceptional. If the two signs are not equal, then it is inferred that the current user's post may have been the cause of the change in the emotional bias of the discussion at that point and the user is added to a list of potentially influential users.

Since our motivation stems from the fact that we are interested in identifying the more influential, for a period of time, users, the ratio alone is not enough and thus, an extra procedure must be followed so as to ascertain their influence on other Twitter users. This sentiment analysis comparison between large numbers of previous and later posts can help us determine in a more accurate way the influence of specific users on other ones. Notably, assume that posts from a concrete user on a discussion prove to have been the reason for alternations in the corresponding discussion emotional bias; then an influence metric value is assigned to this user for this specific discussion.

The rationale behind this last step is that if a user is influential in the current period but had not the same percentage of influence before, then this user is considered to be non-influential for the current discussion. On the other hand, if both present and past discussions of an individual prove to be influential, then this user is added to the list.

4. Implementation

4.1. *Twitter Discussions Synopsis*

Definition 1. A Twitter egg is an account with no followers.

Definition 2. A star is a bipartite graph where the one partition is a singleton. The vertex of this singleton is connected to all remaining vertices.

We implemented the methodology described using Twitter4j[d], a Java based platform utilized for interacting with the Twitter API. The Twitter subgraphs were collected in a time interval of two months, that is (01/06/2017 – 31/07/2017). A topic-based sampling approach was used where tweets are collected via a keyword search query. More specifically, data for three given discussions on Twitter having about 20.000 Tweets, namely #BringBackOurGirls, #HillaryClinton and #Republicans, were downloaded.

The properties of the one out of three datasets (e.g. #BringBackOurGirls) are presented in Table 2. The first column has fundamental graph structure properties such as the number of edges and triangles, whereas the second column has Twitter specific properties such as the average tweet length and the average number of followers. Note that the vertices are accounts and the directed edges represent "following" relationships.

There is a small fraction of eggs and stars. There are 35 stars which are comprised of 190 vertices, namely the 155 eggs plus the 35 stars. The remaining 8971 vertices belong to a single, large component which is densely connected since its diameter equals 9. These structural characteristics indicate an active social network.

As we are interested in creating social media graphs with high degree of dense connections and not just some random graphs, we tried to introduce the same properties regarding the other two datasets.

Table 2. Subgraph properties for #BringBackOurGirls dataset

Property	**Value**	**Property**	**Value**
Vertices	9161	Hashtags	11
Edges	39152	Tweets	21315
Triangles	213	Retweets	9133
Squares	72	Avg. Following	3.17
Stars	35	Avg. Followers	4.13
Components	32	Eggs	155
Diameter	9	Avg. Tweets	102.8

[d] `http://twitter4j.org/en/index.html`

4.2. *Discussions Sentiment Calculations*

Table 3 presents the percentages of sentiment biases on the three corresponding Twitter datasets. What we can observe is that in all discussions, both the three categories are present. Positive polarity describes the emotions that change the emotional stance towards a better situation, while in contrast, negative polarity tends to affect human psychology situation towards an unpleasant direction. Moreover, in political discussions, the negative polarities have lower percentages as users, who have a stable opinion towards a certain person or certain party (e.g. #HillaryClinton and #Republicans), and thus seem to address posts having a positive stance towards them.

On the other hand, the discussion #BringBackOurGirls achieves positive percentage of 68% as users tend to write in a more assertive way in this topic hoping to learn positive news about the 276 female students who were kidnapped from the Government Secondary School in Nigeria.

Table 3. Percentages of Sentiment Bias on 3 Twitter datasets

Discussion	**Pos**	**Neutral**	**Neg**
#BringBackOurGirls	68	9	23
#HillaryClinton	43	39	18
#Republicans	21	56	23

In addition, one major issue is that in the two datasets that are highly correlated to politics, namely #HillaryClinton and #Republicans, the neutral distribution of the examined tweets are 39% and 56% respectively. This is something that is long anticipated since SentiStrength has human-level accuracy for short social web texts in English, except political texts. Thus, this inefficiency of the selected tool could be upgraded by a manual annotation and verification of a subset of the neutral Tweets in the examined domain, i.e. politics.

5. Evaluation

We have conducted our experiments with the use of three out of six Influence Metrics, namely Conversation (Influence Metric 1), Affiliation (Influence Metric 5) and Atomic Influence (Influence Metric 6). The evaluation constitutes of the difference among the Influence Metrics score as a percentage, the Information Diffusion, the identification of the Top-k Users and the Influential Users, based on the area of Sentiment Analysis.

5.1. *Influence Metrics*

As previously mentioned, the Influence Metrics for all users were computed. In order to gain a deeper insight on the variations between the aforementioned Influence

Metrics, we present a figure depicting the difference in percentage among the metrics score as proposed in Subsection 3.3. The results are presented in Figure 2 and are conducted for the top 100 ranked users. The vertical axis represents the percentage of Influence Metrics starting from 100% where horizontal axis represents the 100 users starting from the one with the highest Influence Metric.

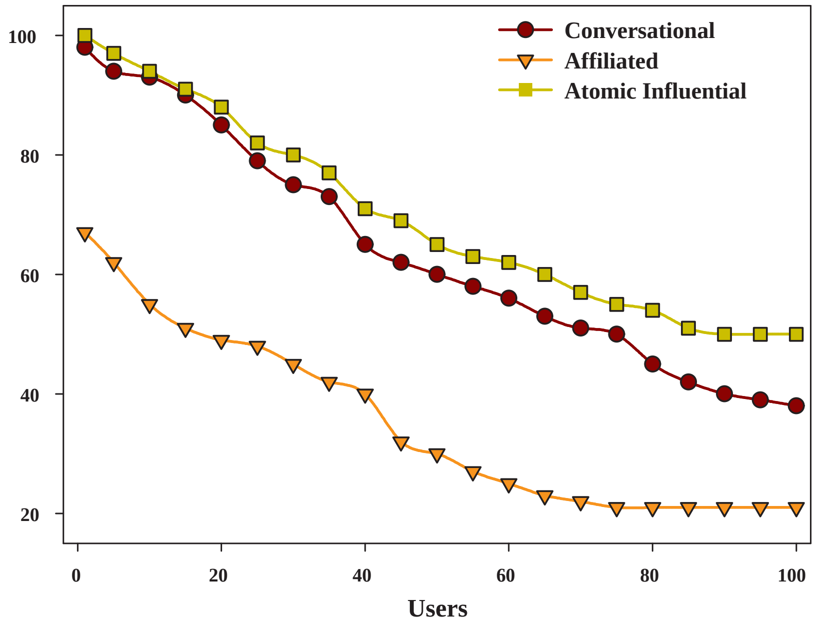

Fig. 2. Influence Metrics Ranking Percentages

It is worth noticing that there are differences in the ranking produced by these three Influence Metrics, as different features are taken into consideration.

The intuition behind this, is to provide the exact metric values for each user who belongs in the higher 100 regarding Influence Metrics and thus, compare these three metrics. Our results show that all Influence Metrics decrease and Conversation as well as Atomic Influence have almost equal values with some variations. Also, both Conversation and Atomic Influence Metrics are stabilized at percentages equal to 40% and 50%, while Affiliation Metric is stabilized at a percentage equal to 20%.

5.2. *Information Diffusion*

Our approach in information diffusion is based on a novel idea that considers fuzzy random graph patterns. We claim that when taking into account user's influence metrics in information diffusion, a more realistic information diffusion process can be constructed. The most important factor which affects the transmission of the tweets is the followers' probability of retweeting [13,27]. So, the percentage of covering a Twitter graph is measured in this subsection.

Figure 3 presents the Rate of Covered Users (or percentage of Graph cover) in case the diffusion starts from the "influential" users. We can therefore observe, that almost the 30% of total users are enough in order to cover the whole graph of users. Hence we can state that our assumption of the "influential" users playing a dominant role in information diffusion, is verified. The vertical axis represents the percentage of Atomic Influence Metric where the user with the highest metric has value equal to 100% and the horizontal axis represents the rate of covered users.

This percentage is derived from the "Following" relationships among users of the generated graph. Specifically, taking as input the ascending order list of the users with the highest Atomic Influence Metric, we observe that with use of the 30% of these users, the users of the whole graph can be informed for a specific post.

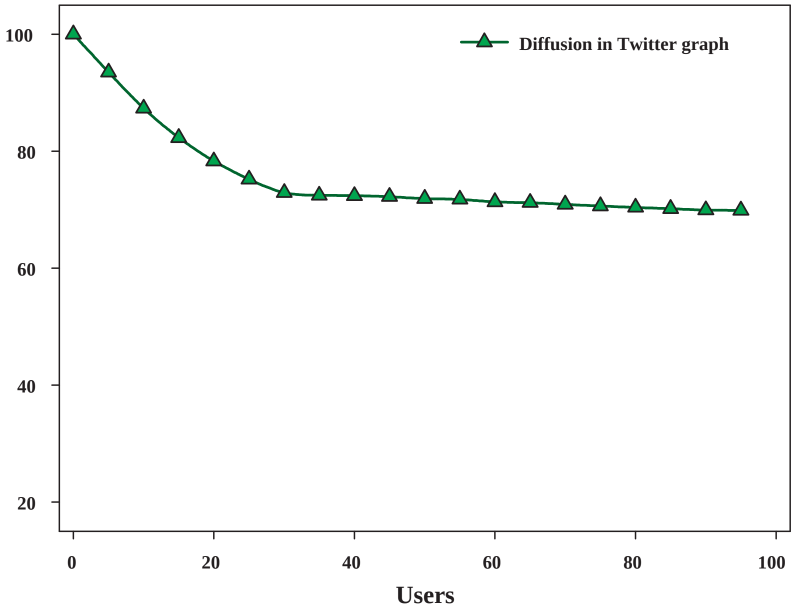

Fig. 3. Rate of Covered Users

5.3. *Top-k Users*

For the purposes of user evaluation of the different result sets, we organized an online survey and asked social media users to anonymously complete some web forms. A web application linked to a database was developed where anonymous answers were concentrated for later process.

A sample of 350 students associated with the University of Patras were given the topics as well as the top-10 users with the highest Influence Metric, and in following they could browse through the tweets. After browsing through the datasets, users were asked to choose the most influential username per topic, according to what they believed. They were presented with three forms, one for each topic, where they were required to rank each of the usernames participating in the topics with a rank between 1 to 10 according to whether they are influential or not. The final rank for a username is the sum of ranks it has gained.

To understand the effectiveness of each proposed Influence Metric, we have used Precision and Pearson Correlation Coefficient metrics [26] in order to measure the correctness of the results and find out whether there is an agreement between the results and user evaluation for the ranking order of users. Figures 4 and 5 present the results related to these two metrics. One can see an improvement in the Precision as well as in the Pearson Correlation Coefficient for the Atomic Influence metric.

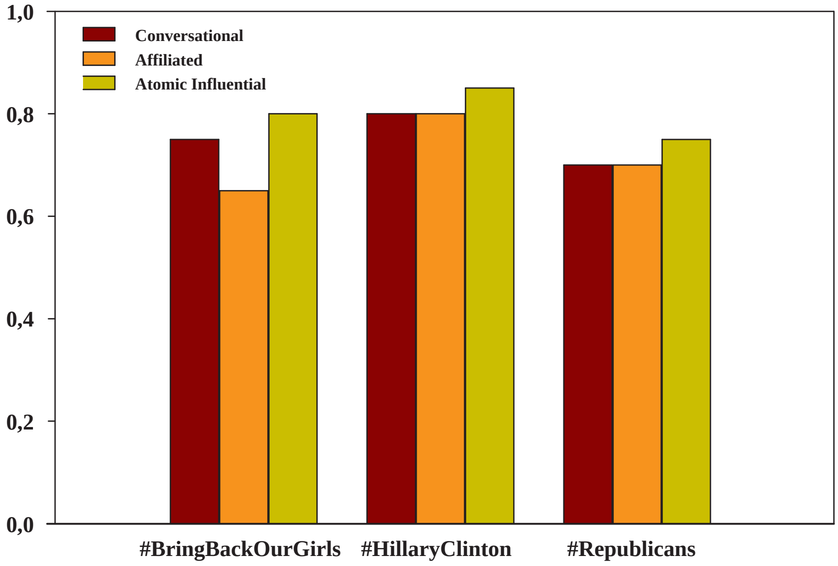

Fig. 4. Precision of each Influence Metric

5.4. *Influential Users*

In order to further evaluate our research proposal, we tried out various combinations of window sizes as stated in Subsection 3.5. To obtain the optimal results, we have used window sizes which do not exceed the 1/3 of the total number of users. This upper limit for the window was selected in order to ensure that user influence is examined in a small scale within the given discussions.

Please notice that the corresponding abbreviations for Table 4 are as follows; Not for Not Influential, Pot for Potentially Influential and Inf for Influential users.

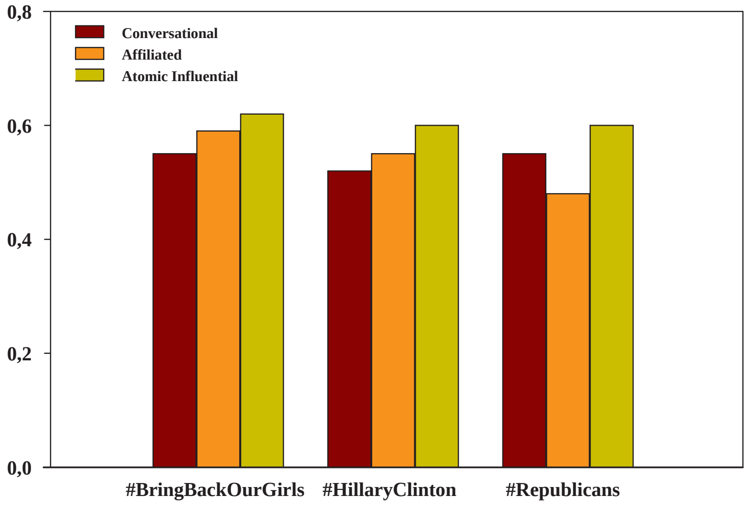

Fig. 5. Pearson Correlation Coefficient of each Influence Metric

From Table 4, it is clear that the percentage of potentially influential users is significantly greater than that of the actually influential ones. Another observation deriving from the results, is that the increase in window size leads to a decrease of the number of both possible as well as actual influential users. By increasing the window size, we include more posts in the input used by the system during the user influence calculation phase of the methodology. In that way, each user's influence on the posts that precede and follow their own, is calculated more accurately.

Table 4. Percentage of Not Influential/Potentially Influential/Influential users on 3 Twitter datasets for 3 different Window Sizes

Hashtag	**Window Size**	**Not**	**Pot**	**Inf**
#BringBackOurGirls	15	72	20	8
#BringBackOurGirls	30	83	13	4
#BringBackOurGirls	60	85	10	5
#HillaryClinton	15	75	20	5
#HillaryClinton	30	81	15	4
#HillaryClinton	60	87	10	3
#Republicans	15	52	36	12
#Republicans	30	77	17	6
#Republicans	60	85	11	4

Finally, the numbers illustrated in Table 4 show that, in most cases, the number of influential users does not exceed the 1/10 of the number of total users in a discussion. This could be coincidental and entirely dependent on the subject of the given discussion. Even if this is the case though, it is clear that influential users do not make posts on any given subject of discussion. The reasons for the existence of this phenomenon seem to be social or psychological.

6. Conclusions

In this paper, we propose a novel method for identifying which of the users, taking part in a given discussion on Twitter, are "influential". More in detail, taking as input data from Twitter topics, we measure several Influence Metrics and in following analyze users' posts on the basis of sentiment bias. By implementing these processes and testing them on Twitter, we discover that the number of influential users tends to be quite small on most discussions, since they seem to avoid making posts on discussions already heavily influenced by other influential users. Another significant contribution constitutes the Fuzzy Random Graph, where each edge between two Twitter users shares an equal probability of contact (via "Following"), thus being independent from the other pairs. In addition, while the diffusion of information in the corresponding graphs was implemented according to the transient contacts model, its spread takes a fair amount of time to evolve.

As work to come, we are interested in predicting the possible future existence of influential users in discussions, which have not been influenced by such users before. Moreover, analysis from dynamic systems' theory or even different algorithmic analytical tools can be incorporated. A more extensive experimental evaluation can additionally be considered as direction for future work. Finally, the adoption of efficient heuristics on time-varying graphs is very promising and thus, can be introduced in our proposed work.

References

1. Lada A. Adamic, Rajan M. Lukose, Amit R. Puniyani, and Bernardo A. Huberman. Search in power-law networks. *Physical Review E*, 64(4):046135, 2001.
2. Ioannis Anagnostopoulos, Vassilis Kolias, and Phivos Mylonas. Socio-semantic query expansion using twitter hashtags. In *7th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pages 29–34, 2012.
3. Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *43rd Hawaii International Conference on Systems Science (HICSS)*, pages 1–10, 2010.
4. Christopher Cannings and David B. Penman. Chapter 2: Models of random graphs and their applications. *Handbook of Statistics*, 21:51–91, 2003.
5. Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. How community feedback shapes user behavior. In *8th International Conference on Weblogs and Social Media (ICWSM)*, 2014.
6. Martin J. Chorley, Gualtiero B. Colombo, Stuart M. Allen, and Roger M. Whitaker.

Human content filtering in twitter: The influence of metadata. *International Journal of Human-Computer Studies*, 74:32–40, 2015.

7. Sergey N. Dorogovtsev. *Lectures on Complex Networks*, volume 24. Oxford University Press, New York, 2010.
8. Georgios Drakopoulos, Andreas Kanavos, Phivos Mylonas, and Spyros Sioutas. Defining and evaluating twitter influence metrics: a higher-order approach in neo4j. *Social Network Analysis and Mining*, 7(1):52:1–52:14, 2017.
9. Paul Erdös and Alfréd Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.
10. Frank A. Haight. *Handbook of the Poisson Distribution.* Wiley, 1967.
11. Paul Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912.
12. Eleanna Kafeza, Andreas Kanavos, Christos Makris, Georgios Pispirigos, and Pantelis Vikatos. T-PCCE: Twitter personality based communicative communities extraction system for big data. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 1(1), 2019.
13. Eleanna Kafeza, Andreas Kanavos, Christos Makris, and Pantelis Vikatos. Predicting information diffusion patterns in twitter. In *10th International Conference on Artificial Intelligence Applications and Innovations (AIAI)*, pages 79–89, 2014.
14. Eleanna Kafeza, Andreas Kanavos, Christos Makris, and Pantelis Vikatos. T-PICE: Twitter personality based influential communities extraction system. In *IEEE International Congress on Big Data*, pages 212–219, 2014.
15. Zhen Kan, John M. Shea, and Warren E. Dixon. Influencing emotional behavior in a social network. In *American Control Conference (ACC)*, pages 4072–4077, 2012.
16. Andreas Kanavos, Isidoros Perikos, Ioannis Hatzilygeroudis, and Athanasios Tsakalidis. Emotional community detection in social networks. *Computers & Electrical Engineering*, 65:449–460, 2018.
17. Andreas Kanavos, Isidoros Perikos, Pantelis Vikatos, Ioannis Hatzilygeroudis, Christos Makris, and Athanasios Tsakalidis. Conversation emotional modeling in social networks. In *26th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 478–484, 2014.
18. Renato Kempter, Valentina Sintsova, Claudiu Cristian Musat, and Pearl Pu. Emotionwatch: Visualizing fine-grained emotions in event-related tweets. In *8th International Conference on Weblogs and Social Media (ICWSM)*, 2014.
19. Hamidreza Keshavarz and Mohammad Saniee Abadeh. Accurate frequency-based lexicon generation for opinion mining. *Journal of Intelligent & Fuzzy Systems*, 33(4):2223–2234, 2017.
20. Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web TWEB*, 1(1), 2007.
21. Chun-Hsien Li, Chiung-Chiou Tsai, and Suh-Yuh Yang. Analysis of epidemic spreading of an sirs model in complex heterogeneous networks. *Communications in Nonlinear Science and Numerical Simulation*, 19(4):1042–1054, 2014.
22. Sin Hong Lin and Ming Hong Liao. Towards publishing social network data with graph anonymization. *Journal of Intelligent & Fuzzy Systems*, 30(1):333–345, 2016.
23. Eduardo Massad, Neli Regina Siqueira Ortega, Cláudio José Struchiner, and Marcelo Nascimento Burattini. Fuzzy epidemics. *Artificial Intelligence in Medicine*, 29(3):241–259, 2003.
24. Cameron Nowzari, Victor M. Preciado, and George J. Pappas. Analysis and control of epidemics: A survey of spreading processes on complex networks. *IEEE Control Systems*, 36(1):26–46, 2016.

25. Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of Modern Physics*, 87:925–979, 2015.
26. David M. W. Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.
27. Gerasimos Razis and Ioannis Anagnostopoulos. Influencetracker: Rating the impact of a twitter account. In *10th International Conference on Artificial Intelligence Applications and Innovations (AIAI)*, pages 184–195, 2014.
28. Gerasimos Razis and Ioannis Anagnostopoulos. Semantifying twitter: the influencetracker ontology. In *9th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pages 98–103, 2014.
29. Kenneth J. Rothman, Sander Greenland, and Timothy L. Lash. *Modern Epidemiology*. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia, 2008.
30. Hongkee Sul, Alan R. Dennis, and Lingyao Ivy Yuan. Trading on twitter: The financial information content of emotion in social media. In *47th Hawaii International Conference on System Sciences (HICSS)*, pages 806–815, 2014.
31. Jared Suttles and Nancy Ide. Distant supervision for emotion classification with discrete binary values. In *14th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, pages 121–136, 2013.
32. Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment in short strength detection informal text. *Journal of the American Society for Information Science and Technology (JASIST)*, 61(12):2544–2558, 2010.
33. Andranik Tumasjan, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *4th International Conference on Weblogs and Social Media (ICWSM)*, pages 178–185, 2010.
34. Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. TwitterRank: Finding topic-sensitive influential twitterers. In *3rd International Conference on Web Search and Web Data Mining (WSDM)*, pages 261–270, 2010.
35. Lotfi A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 100:9–34, 1999.
36. Jiguo Zhang, Gaofeng Liu, and Yanbing Gong. Some notes on characters of intuitionistic fuzzy sets. *Journal of Intelligent & Fuzzy Systems*, 30(2):991–998, 2016.
37. Peiyun Zhang, Rongjian Xie, and Bo Huang. Trustworthy services diffusion based on optimizational nodes in online social network. *Journal of Intelligent & Fuzzy Systems*, 31(4):2281–2290, 2016.