sustainability

MDPI

Article

SentiFlow: An Information Diffusion Process Discovery Based on Topic and Sentiment from Online Social Networks

Berny Carrera and Jae-Yoon Jung *

Department of Industrial and Management Systems Engineering, Kyung Hee University, Yongin-si, Gyeonggi-do 446-701, Korea; berny@khu.ac.kr

* Correspondence: jyjung@khu.ac.kr; Tel.: +82-31-201-2537

Received: 3 July 2018; Accepted: 31 July 2018; Published: 2 August 2018

Abstract: In this digital era, people can become more interconnected as information spreads easily and quickly through online social media. The rapid growth of the social network services (SNS) increases the need for better methodologies for comprehending the semantics among the SNS users. This need motivated the proposal of a novel framework for understanding information diffusion process and the semantics of user comments, called SentiFlow. In this paper, we present a probabilistic approach to discover an information diffusion process based on an extended hidden Markov model (HMM) by analyzing the users and comments from posts on social media. A probabilistic dissemination of information among user communities is reflected after discovering topics and sentiments from the user comments. Specifically, the proposed method makes the groups of users based on their interaction on social networks using Louvain modularity from SNS logs. User comments are then analyzed to find different sentiments toward a subject such as news in social networks. Moreover, the proposed method is based on the latent Dirichlet allocation for topic discovery and the naïve Bayes classifier for sentiment analysis. Finally, an example using Facebook data demonstrates the practical value of SentiFlow in real world applications.

Keywords: information diffusion; community detection; topic analysis; sentiment analysis; social networks

1. Introduction

Today, social network services (SNS) are an effective medium through which new information, such as opinions, news, and advertisements, is easily and quickly disseminated. The spread of these information starts when users create new posts. Subsequently, all subscribers and users who comment are notified of the new posts. To better understand the spread of these ideas, it is important to analyze how people propagate their thoughts based on their opinions and topics of interest, which are the underlying context and information flow.

Some applications for the proposed technique are viral marketing, where marketers quantify the impact of released products by applying sentiment analysis to understand unsatisfied consumers; influence analysis, determining how groups of users influence other groups; and trend detection, in which with the application of topic models, discussions, and opinions can be uncovered. In particular, in terms of social science, it is possible to understand the relationship among the users' behaviors, the distinctions of communities and the information diffusion in social networks. Understanding users' reactions are valuable since opinions can influence the news trend or purchase decisions. Therefore, the users' opinions are vital to understanding the way information spreads and how communities interact among them.

There have been several studies on information flow modeling based on the structure of social network or the discovery of information diffusion processes without analyzing the structure of

communication [1–6]. However, their studies have not shown the relation between context and information flow. A few studies have considered how to model information diffusion with a process structure. Kim et al. [7] presented an information diffusion model using data of a blog to analyze the reposting behaviors of people. Although other studies have been carried out on information flows and SNS data [8–11], they can be used only to infer an information diffusion flow without considering the probability that communities will communicate or showing the contextual information. Kim et al. [12] analyzed the behavioral patterns in SNS, News and Blog sites. However, the disadvantage in their study is the absence of opinions in the users' comments. Opinions made by users in SNS can have a huge impact in society, therefore the analysis of these emotions are important to monitor the response of users to a specific news or product [13]. There exist some studies in information diffusion based on sentiment [14–16], but they do not consider the opinion flow between communities. Other researchers analyzed how to visualize topics and opinions in SNS [17–21], but they lack in the information process flow. In this research, a new semantic hidden Markov model (HMM) for discovering information diffusion, named SentiFlow, is introduced to discover probabilistic information flow in consideration of topics and sentiment. It is an extension of HMM [22] using text mining and process mining [23]. The probabilities in the SentiFlow are computed based on maximum likelihood (ML) [22]. In our previous studies [24,25], a method for probabilistic information flow of the communication between users and communities is presented. A method to underline the semantics and opinions in the interactions among user groups is suggested in this paper by applying community clustering algorithms to find user communities and by undertaking two different analyses, topic modeling and sentiment analysis, for the user comments. Finally, the traces of these communications are analyzed, and different information flow process models are generated. The goal is to answer the following important questions: (1) "What topics promote communication among user communities?" (2) "How are the positive, neutral, and negative opinions shared in the information diffusion process from a probabilistic point of view?"

The rest of the paper is organized as follows. Section 2 describes the proposed methodology used in this work. Section 3 describes the general algorithm used. Section 4 provides the experimental results. Finally, Section 5 concludes this work.

2. Framework

In this research, log data collected from SNS are used to discover information diffusion process. Generally, users in SNS wrote posts and their friends comment on the posts. Therefore, it can be assumed that, in a SNS log, each post of a specific user is characterized by many comments of his/her friends and the comments are ordered chronologically. From the SNS log, we first find the user communities based on their interaction and draw the information diffusion process. The topics of interest are found and annotated on the discovered process. Finally, the sentiment for the topics is analyzed for topics and users. The overall framework is depicted in Figure 1.

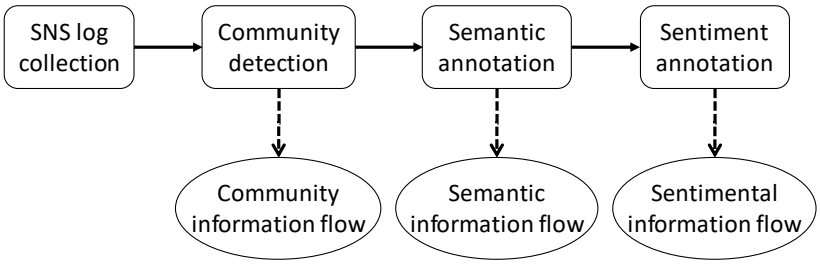

Figure 1. A framework of the information diffusion process discovery with topic and sentiment.

2.1. SNS Log Collection

In SNS such as Facebook or Twitter, all posts are obtained along with the user's name, user's comments, and the timestamp indicating when it was posted to create an SNS log. In this research, it is assumed that users write comments to reply to other users and thus create or continue a discussion about a related post. Each post published by the fan page owner is characterized with a sequence of SNS events. An SNS event contains a user, the user comment, and the time when the comment was published. This sequence is ordered using the timestamp of the comment publication. An SNS log is defined below.

Definition 1. *(SNS log) Let* $P = \{p_1, \ldots, p_K\}$ *and* $U = \{u_1, \ldots, u_V\}$ *be the finite sets of all possible post identifiers and users, respectively. K and V are the numbers of posts and users, respectively. Posts are characterized by SNS events e, which in turn are characterized by various attributes att. For any SNS event* $e \in E$, $\#att(e)$ *is the value of attribute att in event e to have SNS event* $e = (\#u(e), \#user\_comment(e), \#time(e))$. *Additionally, each post has an attribute action trace for a specific post p, denoted by* σ_k, *and is defined as the sequence of SNS events in p, i.e.,* $\sigma_k = \langle e_1, \ldots, e_H \rangle$ *for* $1 \leq k \leq K$, *where H is the number of events for p. An SNS log, denoted by* $L = [\sigma_k]$, *is a multi-set of action traces over U and P in the SNS.*

To illustrate the operation of the proposed framework, Table 1 presents an example SNS log with synthetic data. The example SNS log contains the post identification and the comment traces. The comment traces show the structure User $_{\text{comment}}$, where the user's name is written and followed by the comment in subscript. Each user and comment are ordered by the timestamp for when the comment was published. The example SNS log contains six posts and 23 comments written by five users: Angela, George, John, Paul, and Ringo.

Table 1. An example SNS log L_1. A SNS log contains many action traces, which are sequential comments replies to specific posts.

Post ID	Action Trace
p_1	John $_{\text{I like it}}$, Angela $_{\text{This is amazing!}}$, George $_{\text{I think this is absurd}}$
p_2	John $_{\text{We need to be persistent}}$, Ringo $_{\text{I think this is very aggressive}}$, Paul $_{\text{I am ashamed}}$, George $_{\text{We need to demand our rights!}}$
p_3	John $_{\text{It's better if we reform the laws}}$, Paul $_{\text{I am relaxed}}$, Ringo $_{\text{This is a revolution}}$, George $_{\text{pitiful}}$
p_4	John $_{\text{Wow this is perfect}}$, Paul $_{\text{That is bad}}$, Ringo $_{\text{Nice}}$, George $_{\text{Excellent}}$
p_5	John $_{\text{Too much stress}}$, Paul $_{\text{I am afraid}}$, Ringo $_{\text{Superficial}}$, George $_{\text{It is ok everything will be fine}}$
p_6	John $_{\text{Terrorism}}$, Ringo $_{\text{I am so tired}}$, Paul $_{\text{I am so happy}}$, George $_{\text{Cool}}$

2.2. Information Flow among Communities

In this step, the communication of users and the interaction between them are analyzed. For this, users with similar behavior can be clustered into communities. The community detection analysis performs the next activities: identify the network structure inside the SNS log by applying community-detection algorithms, determine how the people across the comments are related, and help minimize the complexity of the discovered process model. The discovered communities represent the community states in the process model.

In the proposed framework, the Louvain modularity (LM) algorithm is used, which is often applied as a community detection method in social network analysis [26]. LM detects and extracts communities in a network by providing the optimal number of communities and optimizing the value of modularity [26], the results of which is used for the best grouping of users in this research. The user communities for this research also represent the information diffusion states for the process model. Moreover, the objective in this step is achieved by creating the information diffusion matrix. The information diffusion matrix represents the frequency of communication inside, outside, and among the communities. Thereby, the action traces in the posts can represent the sharing of information between user communities, which creates an information diffusion matrix to represent the

information flow frequencies from one community to another. A process model is then obtained as output. The user community and the diffusion community matrix are defined as follows.

Definition 2. *(User community) Let U be a finite set of users in SNS log L and* $C = \{c_1, \ldots, c_N\}$ *for* $1 \leq i \leq N$ *be a finite set of communities of users in L. A user community* $c_i \subseteq U$ *is a subset of users grouped by the results of a community detection algorithm* $community(u)$.

Definition 3. *(Information diffusion matrix) Let C be user communities of SNS log L. The information diffusion matrix contains the information flow frequencies between two communities in C, which is denoted by* $A = (a_{ij})$, *where* $a_{ij} = \sum_{\forall \sigma \epsilon L} |c_i \rightarrow c_j|$ *represents the sum of the frequencies of information diffusion from* c_i *to* c_j *and* $\pi = \{c_1, \ldots, c_N\}$, *where* π_i *is the probability of being in* c_i *at time 1 in every action trace* $\sigma \epsilon L$.

$$a_{ij}{}' = \frac{a_{ij}}{\sum_{n=1}^{N} a_{in}} \tag{1}$$

The information diffusion process model indicates the beginning of the information diffusion and how the information spreads among the communities. The model mines the initial probability π that describes the probability of which user community starts the information diffusion in each action trace. The calculation of the parameter values using A for the process model is shown in Equation (1). $a_{i,j}{}'$ is obtained from the information flow frequency of the community c_j, which follows community c_i, divided by the total information flow frequency of all communities that follow c_i.

From the example introduced in Table 1, it is seen that the communities with LM, for this example the resolution parameter value = 0.02, is used to show a better structure of the information diffusion from smaller clusters. The results of the LM algorithm are c_1 = {John, Angela}, c_2 = {George, Ringo}, and c_3 = {Paul}. Table 2 presents the first findings for the process model obtaining the matrix diffusion community A and the probability of state transition A'. In addition, the transition probability distribution is π = (1, 0, 0) where c_1 always initiates the information diffusion.

Table 2. Matrices extracted from L_1: (a) information diffusion matrix A; and (b) state transition probability matrix A'.

	(a)			(b)		
	c_1	c_2	c_3	c_1	c_2	c_3
c_1	0	3	3	0.0	0.5	0.5
c_2	0	0	2	0.0	0.0	1.0
c_3	0	5	0	0.0	1.0	0.0

The information diffusion process model can be drawn in Figure 2. It is clear that the thickness of c_2 is greater because it has more incoming information diffusion than do the other communities. Moreover, the information diffusion between communities c_2 and c_3 is notable.

The information flow in this research is based on the detected communities as shown in Figure 2. However, the communities may be changing over time. To obtain more reliable structure of the information flow, data in enough long period is needed for community detection.

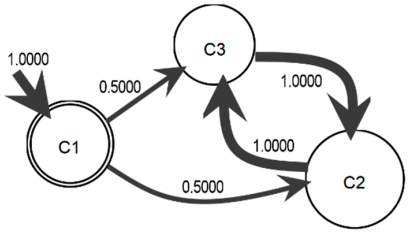

Figure 2. The information diffusion process model generated from log L_1.

2.3. Semantic Information Flow

The third step is the discovery of underlying topics where the meaning and sense of data are analyzed by extracting the principal keywords from comments to understand what people's interests are and how the topics relate to them. As input, the user comments are collected, and tokenization of the words is conducted by breaking the comments into sentences and then into tokens to remove the English non-words, punctuation, and stop words. In this study, the probabilistic topic model technique called the latent Dirichlet allocation (LDA) is used to properly assign and discover the hidden context from the data. LDA can represent documents, signified by the comments of users as mixtures of topics, and then assigns the words with certain probabilities [27]. Furthermore, in the semantic information diffusion process model, the frequent topics inside a community and the probability of those topics are analyzed.

To construct this semantic process model, analysis of the comments needs to be completed as described above. The first goal of this step is to find a small number of topics from the observations of user comments. Hence, the LDA algorithm is used to discover these topics. Thus, each topic found is assigned to each comment to finally obtain the topic matrix as a second goal. The topic matrix represents the frequency with which each user from a community publishes a comment for a specific topic. Therefore, the topic matrix can be defined.

Definition 4. *(Topic matrix) Let C be information diffusion states of the information diffusion process of SNS log L and $T = \{t_1, \ldots, t_M\}$, with $1 \leq m \leq M$ being a finite set of topics discovered from the LDA algorithm $lda(\#user\_comment(L))$. The topic matrix is denoted by $B = (b_{im})$, where each element $b_{im} = \sum_{\forall \sigma \epsilon L} f(c_i, t_m)$ contains the sum of the frequencies in which a topic t_m exists in the user comments of a community c_i in every action trace $\sigma \epsilon L$.*

$$b_{im}{}' = \frac{b_{im}}{\sum_{q=1}^{M} b_{iq}} \tag{2}$$

The calculation of the parameter values using B for the information diffusion process model is shown in Equation (2). $b_{im}{}'$ is obtained from the frequency of community c_i, which is paired with a topic t_m, divided by the total frequency of all topics paired with c_i.

This step uses the community states, topic observations, and SNS log as inputs and generates a semantic process model as output. Moreover, a topic matrix is constructed and shows the frequency between the observed context and each community. The topics are determined from the user's comments. To discover the topics, the LDA algorithm is used, and two topics are obtained from the comments of L_1. The top eight keywords of the two topics are t_1 = {need, nice, demand, happy, need demand, need persistent, ok, everything fine} and t_2 = {think, wow perfect, pitiful, better, better reform, cool, everything, excellent}.

Table 3. Matrices extracted from log L_1 for constructing a semantic information diffusion process model: (a) topic matrix B; and (b) observation symbol probability matrix B'.

	(a)		(b)	
	t_1	t_2	t_1	t_2
c_1	4	3	0.57	0.43
c_2	3	8	0.27	0.73
c_3	4	1	0.80	0.20

Table 3 presents the first findings for the semantic process model obtaining the topic matrix B and the observation symbol probability matrix B'. Figure 3 shows the semantic information flow drawn from B'. In the graphical representation, the dashed arcs created from the community state of the topic represent the interest of the community in the specific topic, and the thickness of the arc represents the probability between them.

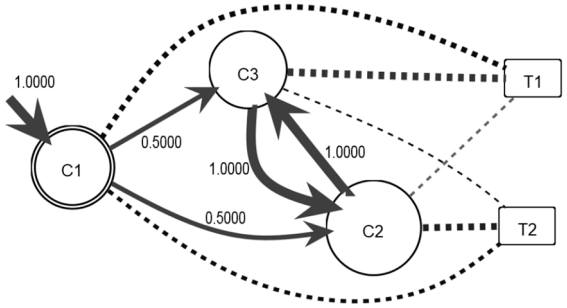

Figure 3. The semantic information diffusion process model generated from log L_1. t_1 and t_2 are the topics that were discovered from the texts exchanged among users.

2.4. Sentimental Information Flow

In this step, sentiment analysis, which is classification of the user comments based on sentiment, is performed to gain a better understanding of the user. The analysis is performed using the naïve Bayes sentiment classifier described in [28] because this method shows good performance in many applications [29]. In this research, the polarity of the comment is classified as a positive, neutral, or negative impression of the topics on which the users have commented.

The last output is the sentimental information flow. The model uses the previous matrices and creates a sentiment matrix that represents the probabilities of the sentiments of each community for a specific topic. The sentiment matrix is described below.

Definition 5. *(Sentiment matrix) For SNS log L, let C be a set of user communities in L, T be a finite set of topics in L, and $S = \langle s_1, s_2, s_3 \rangle$ be a tuple of positive, neutral, and negative sentiments discovered from naïve Bayes sentiment classifier $nbsc(\#user\_comment(e))$. The sentiment matrix is a three-dimensional matrix $D = (d_{imr})$, where $d_{i,m,r} = \sum_{\forall \sigma \epsilon L} f(c_i, t_m, s_r)$ is the sum of the frequencies in which a sentiment s_r exists in the user comments of a community c_i for a topic t_m in every action trace $\sigma \epsilon L$.*

$$d_{imr}{}' = \frac{d_{imr}}{\sum_{q=1}^{3} d_{imq}} \tag{3}$$

The calculation of the parameter values using D for the sentimental information flow model is shown in Equation (3). $d_{imr}{}'$ is obtained from the frequency of community c_i and topic t_m, that is paired with a sentiment s_r, divided by the total frequency of all sentiments paired with c_i and t_m. After the construction of the diffusion community matrix, topic matrix, and sentiment matrix, the last step is the modeling of the sentimental information flow, called SentiFlow. A SentiFlow model can be defined as follows.

Definition 6. *(SentiFlow) A SentiFlow model of SNS log L is an extension of HMM for representing semantic and sentimental information diffusion. A SentiFlow model is denoted by $\Lambda(L) = (\pi, C, T, A', B', D')$, where π is the transition probability distribution of initial states, C is a set of user communities, T is a set of discovered topics, A' is the matrix of state transition probability distribution from information diffusion matrix A, B' is the matrix of observation symbol probability distribution from topic matrix B, and D' is the three-dimensional matrix of sentiment probability distribution from sentiment matrix D. Note that $\sum_j a'_{ij} = 1$ for $\forall i$, $\sum_m b'_{im} = 1$ for $\forall i$, and $\sum_r d'_{imr} = 1$ for $\forall (i, m)$.*

$\pi = (\pi_i)$ for $1 \leq i \leq N$, where π_i is the probability of being in c_i at time 1.
$C = \{c_1, \ldots, c_N\} \in L$ for $1 \leq i \leq N$, where $\{c_1, \ldots, c_N\}$ are the information diffusion states of L.
$T = \{t_1, \ldots, t_M\} \in L$ for $1 \leq m \leq M$, where $\{t_1, \ldots, t_M\}$ are the observed topics of L.

$A' = (a'_{ij})$ for $1 \leq i, j \leq N$, where a'_{ij} is the probability of state transition from c_i to c_j.
$B' = (b'_{im})$ for $1 \leq i \leq N, 1 \leq m \leq M$, where b'_{im} is the probability of observing t_m in state c_i.
$D' = (d'_{imr})$ for $1 \leq i \leq N, 1 \leq m \leq M, 1 \leq r \leq 3$, where d'_{imr} is the probability of observing a sentiment s_r from a topic t_m in a state c_n.

Figure 4 provides the representational model with the notation used in this research. It should be noted that the probabilities of B' are shown, but simply indicate the thickness of the arc from a community to the respective topic.

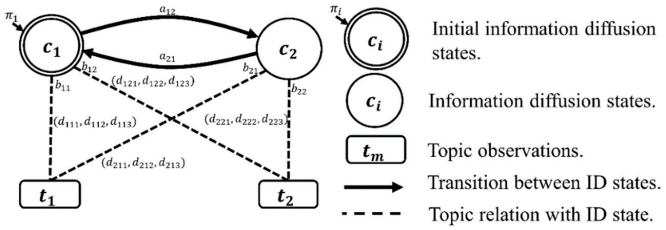

Figure 4. Representation of a SentiFlow model.

The sentimental analysis step in our framework is for understanding the user opinions written in the comments and obtaining a sentiment matrix that has the frequencies from three types of sentiments (positive, negative, and neutral) from the discovered topics in the semantic annotation step. For the sentiment analysis, the naïve Bayes sentiment classifier is used to classify 12 positive, 10 negative, and 1 neutral commentaries. Table 4 presents the findings for the sentimental annotation by obtaining a sentiment matrix D and a sentiment probability matrix D'.

Table 4. Matrices extracted from log L_1 for sentiment annotation: (a) sentiment matrix D; and (b) sentiment probability matrix D'.

		(a)			(b)		
		s_1	s_2	s_3	s_1	s_2	s_3
c_1	t_1	3	0	1	0.75	0.00	0.25
	t_2	2	0	1	0.67	0.00	0.33
c_2	t_1	1	1	1	0.33	0.33	0.33
	t_2	4	0	4	0.50	0.00	0.50
c_3	t_1	1	0	3	0.25	0.00	0.75

Figure 5 presents the graphical representation of the SentiFlow model constructed from log L_1. The difference of color between the communities and topics is shown. For example, c_1 to t_1 shows a bluish color representing predominant positive commentaries (0.75) compared to negative commentaries (0.25). However, c_3 to t_1 presents predominantly negative commentaries for t_1 (0.75), with 0.25 positive commentaries. Additionally, community c_2 to topic t_2 has a mixture of sentiments in the comments, with 0.5 for both. The mapping color of the arc from a community to a topic represents the type of sentiment; the arc is red if the sentiment is negative, lime if neutral, and blue if positive.

A SentiFlow model provides the required information to answer the two questions presented at the end of Section 1. The first question is about the topics that promote communication between communities. In this example, the communication between communities c_2 and c_3 was about topics t_1 and t_2, although c_2 mainly focused on t_2 and c_3 mainly focused on t_1. The second question is about how the sentiment is shared in the information diffusion process from a probabilistic perspective. As an example, the information diffusion from communities c_1 to c_2 for topic t_2 is used. Considering the sequence, <positive, positive>, the result can be analyzed using the forward algorithm [22], and the probability of the sequence is P (<positive, positive> | Λ) = 1.0 × 0.67 × 0.5 × 0.5 = 0.1675. In the case of the sequence <neutral, neutral>, the probability is 0, and the sequence <negative, negative> is

$P(<\text{negative, negative}> \mid \Lambda) = 1.0 \times 0.33 \times 0.5 \times 0.5 = 0.0825$. Therefore, the probability that community c_2 responds positively to a positive comment of community c_1 is higher because community c_1 has a higher probability of posting a positive comment.

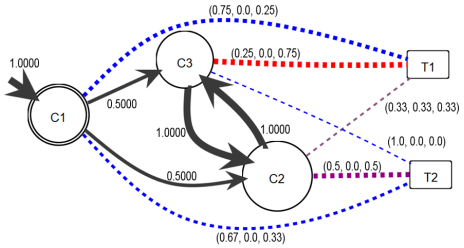

Figure 5. The SentiFlow model generated from log L_1.

3. Algorithm

In this section, the overall procedure that was introduced in the proposed framework is described as an algorithm. The SentiFlow algorithm creates the structure $\Lambda(L) = (\pi, C, T, A', B', D')$ from an SNS log $L = [\sigma]$ similar to an HMM structure $\lambda(L) = (\pi, States, Observations, A', B')$, as shown in Algorithm 1. In the algorithm, LM is adopted for community algorithm detection, and it starts with the clustering of users U in the log (Lines 2–3). Next, for all traces in the log, the algorithm discovers from the user comments, first, the topics T as a result of the LDA algorithm $lda\left(\#user_{comment(e)}\right)$ and, second, the classification of the sentiments S from the naïve Bayes sentiment classifier $nbsc(\#user\_comment(e))$ for each user comment (Lines 4–6). Then, for each trace in the log, the algorithm finds initial communities π, diffusion community matrix a_{ij}, topic matrix b_{im}, and sentiment matrix d_{imr} (Lines 8–25). In particular, if two adjacent users belong to the same community, the algorithm skips the count in the diffusion community matrix, and the last SNS event is counted for its topic and sentiment. Afterwards, the state transition probability matrix A', the observation symbol probability B', and the opinion probability matrix D' are calculated from A, B, and D using ML. Finally, the algorithm returns a SentiFlow model, $\Lambda(L) = (\pi, C, T, A', B', D')$.

Algorithm 1. SentiFlow

1: **Input:** SNS log $L = [\sigma]$, which is a multi-set of action traces σ in the SNS.
2: **Output:** A SentiFlow model, $\Lambda(L) = (\pi, C, T, A', B', D')$
3: Insert all users in L into a user set U.
4: Detect communities C from users U, and prepare function $c = community(\#u(e))$.
5: **For** each trace $\sigma = \langle e_1, \ldots, e_H \rangle$ in L **Do**
Discover topics T from user comment, and prepare a function
6: $t = lda(\#user\_comment(e))$.
7: Discover sentiments S, and prepare a function $s = nbsc(\#user\_comment(e))$.
8: **End For**
9: **For** each trace $\sigma = \langle e_1, \ldots, e_H \rangle$ in L **Do**
10: **If** e_1 **Then**
11: Increase π_i in $community(\#u(e_1))$.
12: **End If**
13: **For** each adjacent SNS event (e_h, e_{h+1}) in σ for $1 \leq h \leq H - 1$ **Do**
14: $c_i = community(\#u(e_h))$ and $c_j = community(\#u(e_{h+1}))$.
15: $t_m = lda(\#user\_comment(e_h))$ and $s_r = nbsc(\#user\_comment(e_h))$.

If $c_i \neq c_j$ **Then**
Increase a_{ij} in A by 1.
End If
Increase b_{im} in B by 1.
Increase d_{imr} in D by 1.
If $e_{h+2} = null$ **Then**
Increase b_{jm} in B by m and $t_m = lda(\#user\_comment(e_{h+1}))$.
Increase d_{jmr} in D by m,r and $s_r = nbsc(\#\text{user}\_comment(e_{h+1}))$.
End If
End For
End For
Calculate the state transition probability matrix $A' = (a'_{ij})$ based on $A = (a_{ij})$.
Calculate the observation symbol probability matrix $B' = (b'_{im})$ based on topic matrix $B = (b_{im})$.
Calculate the sentiment probability matrix $D' = (d'_{imr})$ based on the sentiment matrix $D = (d_{imr})$.
Return a SentiFlow model, $\Lambda(L) = (\pi, C, T, A', B', D')$

4. Experiments

In this research, the SentiFlow algorithm was implemented as a plug-in of the ProM platform to verify the proposed framework. ProM is the open source platform that provides practical applications for process mining and supports many kinds of process discovery algorithms [23].

To illustrate the proposed algorithm, the posts of the CNN Facebook page from 1–5 April 2017 were used. The data contain 208 posts with a total of 67,831 users participating with 143,876 comments from 1 April to 6 June 2017.

To obtain information flow among communities, the community detection was analyzed by applying the LM algorithm. Then, the data were filtered to reduce the noise generated by the infrequent users; as a result, six communities were detected using a resolution parameter of 0.8 [26]. The six detected communities, c_1 to c_6, contain 203, 1048, 13, 9, 25, and 121 users, respectively, among a total of 1419 users.

The result of the information diffusion process discovery based on detected communities is shown in Figure 6; the number of comments in a community is represented by the size of the corresponding node in the figure, and the thickness of an arrow denotes the probability of information diffusion from one community to another. Community c_2 concentrates most of the information flows from c_1, c_3, c_4, c_5, and c_6, revealing a larger size from the higher incoming information flow from smaller communities and the number of user comments. Moreover, the information flow received from c_3 to c_2 shows the highest information diffusion probability among all communities. The threshold of information diffusion probability used for the process model visualization in Figure 6 is 0.04. The threshold is used to present a readable process model removing the arcs with lower probability.

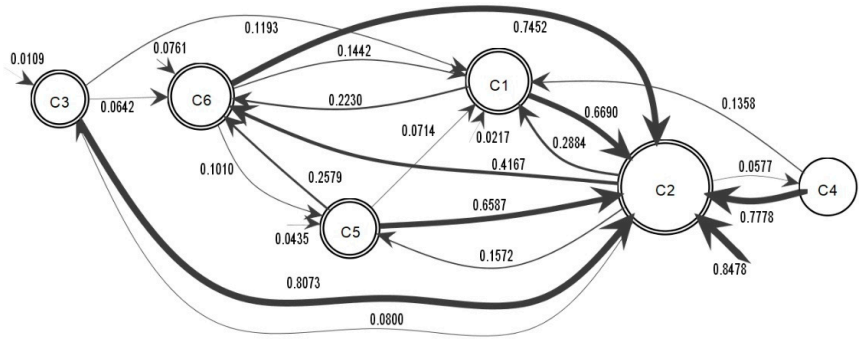

Figure 6. The information flow among communities generated from the CNN Facebook page.

The topic annotation step started with analysis of the comments. First, the stop words, English non-words, and punctuation were removed. Second, duplicate and empty comments were removed. As a result, 13,706 comments and 92 posts were evaluated. To find the different topics of the comments, each word was tokenized as an input for the LDA algorithm. Figure 7 presents a cloud word visualization of the token results for user comments.

Table 5 presents the five topics discovered from the LDA algorithm with their top eight keywords from the discovered comment topics. As shown, topics t_2 and t_3 share two keywords. The word "Trump" is repeated in t_1, t_2, t_3, and t_4 with notable importance in a mixture of topics, but relays in categorize individually each topic.

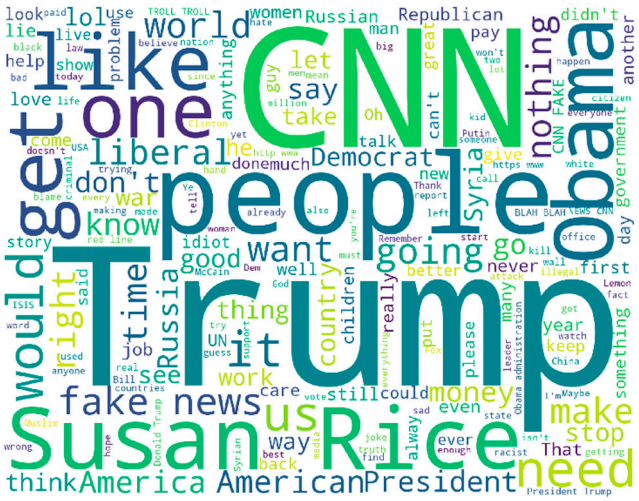

Figure 7. Word cloud of comment keywords of the CNN Facebook page.

Table 5. Top 8 keywords discovered by the LDA algorithm.

Topic	Top 8 Keywords
t_1	money, Trump, troll, pay, make, blah, need, wall
t_2	people, like, get, would, Trump, one, go, women
t_3	Trump, Obama, president, war, Syria, world, people, us
t_4	rice, Susan, Trump, Susan Rice, CNN, Obama, Russia, story
t_5	CNN, news, fake, fake news, Fox, Clinton, lol, lemon

Figure 8 describes illustrates the semantic information flow between user communities. Here, the discovered topics from the LDA algorithm are shown as rectangles along with their identification name. The dashed arcs indicate the use of the topics from the communities. The topic t_2 has greater importance because it has many thicker arcs connecting communities than do other topics. Communities c_1, c_3, and c_4 present frequent use of keywords for topics t_2 and t_3. As in the previous step, the threshold used to present the information flow is 0.04.

The last step is the generation of the sentimental information flow shown in Figure 9. The model describes the probability of opinions by drawing the arc to a positive community in blue, neutral in green, and negative in red. In more detail, a label with three probabilities of positive, neutral, and negative comments is added on the corresponding arc in order. An example of a negative opinion is shown by a reddish dashed arc representing c_3 over t_2. Conversely, a positive probability opinion can be observed from c_3 toward t_3 with a bluish color. The figure shows that c_5 toward t_2 shows a mixture of opinions and has relative balance between positive and negative opinions. In general, neutral opinions show a lower probability than positive and negative opinions.

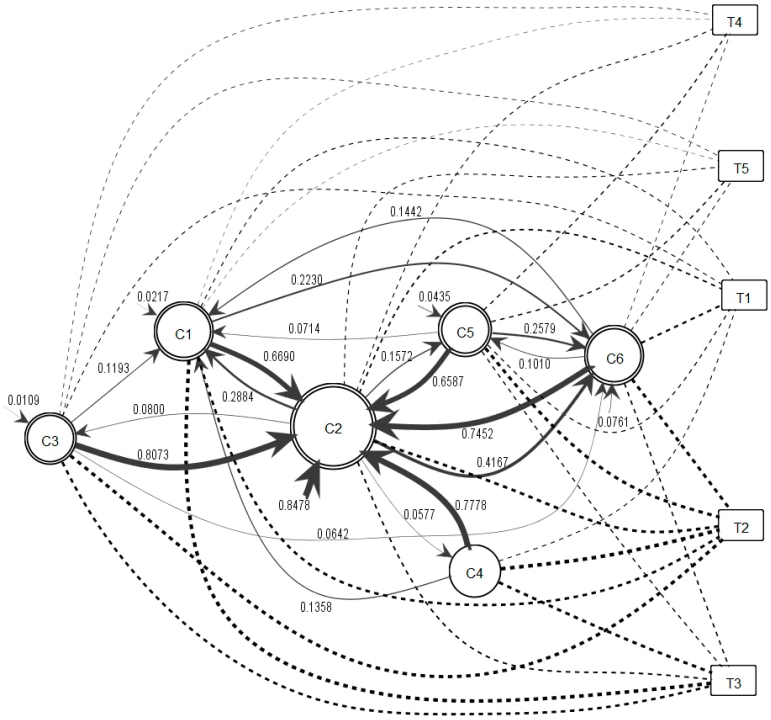

Figure 8. The semantic information flow generated from the CNN Facebook page.

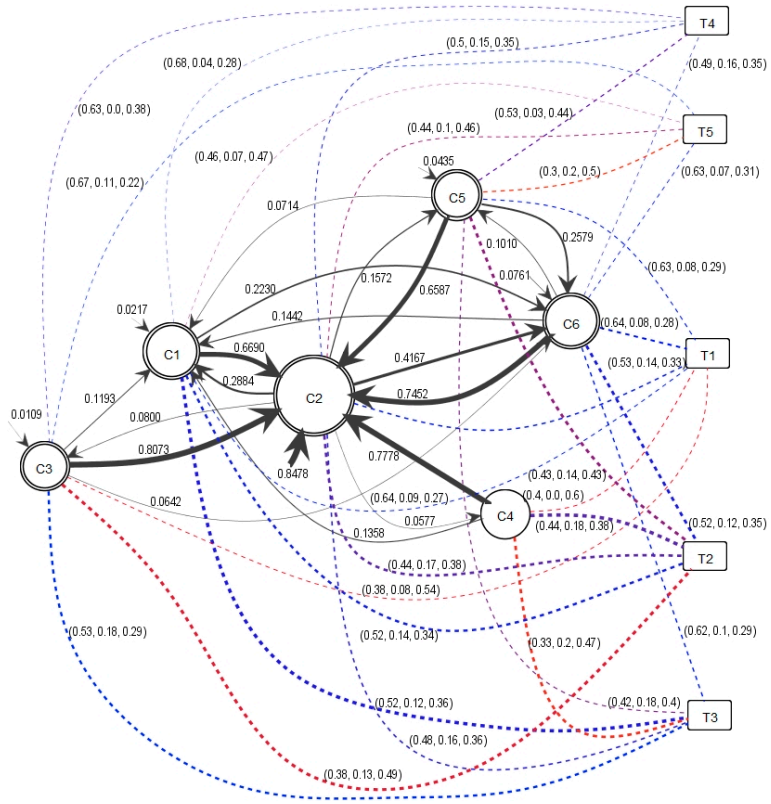

Figure 9. The sentimental information flow based on polarity for topics generated from the CNN Facebook page.

Figure 10 illustrates the relationship between the five discovered topics and the opinions with a total of 6820 positive, 4957 negative, and 1910 neutral comments, separated into the six communities. There are two trend topics, t_2 and t_3, followed by t_1 and t_4 and finally t_5 as the least discussed. In addition, community c_5 has a similar amount of interest between topics t_1, t_3, t_4, and t_5. Furthermore, topic t_2 is the most commented upon among all the communities with the exception of community c_1, which focuses on topic t_3.

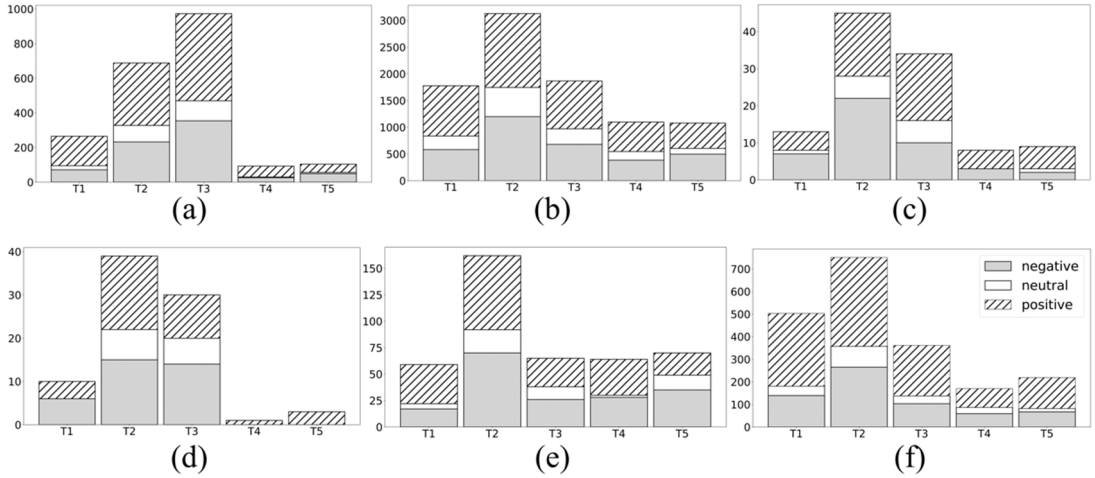

Figure 10. Sentiment analysis across user opinions by each topic among six communities: (**a**) community c_1; (**b**) community c_2; (**c**) community c_3; (**d**) community c_4; (**e**) community c_5; and (**f**) community c_6.

The individual information flows for each topic are shown in Figure 11 using the threshold of 0.04. In the different information flows, community c_2 is continuously the largest community and concentrates most of the information flows from c_1, c_3, c_4, c_5, and c_6 from the different topics. Figure 11a shows a SentiFlow model from topic t_1 with a different flow from Figure 9, where community c_3 does not provide an initial probability and indicates an information diffusion to c_4 with probability 0.0769. Additionally, c_3 and c_4 have a predominant negative opinion in contrast to c_1, c_2, c_5, and c_6 with a positive opinion. In Figure 11b, the initial information diffusion π changed for community c_4 not presented in other SentiFlow models with probability of 0.0110. In addition, c_4, c_5, and c_2 have a purple color to note they have an opinion divided between negative and positive. However, c_6 shows a greater positive opinion, whereas c_3 presents a greater negative opinion. In Figure 10a, community c_1 has the most comments for topic t_3, but, in Figure 11c, c_1 is smaller than c_2 because there are more user comments than in c_1. Positive opinions are expressed in c_1, c_2, c_3, and c_6, whereas while negative opinions are expressed in c_4, and mixed opinions are expressed by users from c_5. In Figure 11d,e, good information flow is observed between all communities with the exception of c_4. In this case, the community does not show an incoming information flow because the probabilities are below 0.04. For Figure 11d, a general predominant positive opinion can be seen for almost all communities, even though c_5 has a combination of positive and negative opinions. In the sentiment information diffusion for topic t_5, c_3, c_4, and c_6, have a positive opinion, in contrast to c_5 with negative comments and c_1 and c_2 with a balanced opinion between positive and negative, as shown in Figure 11e.

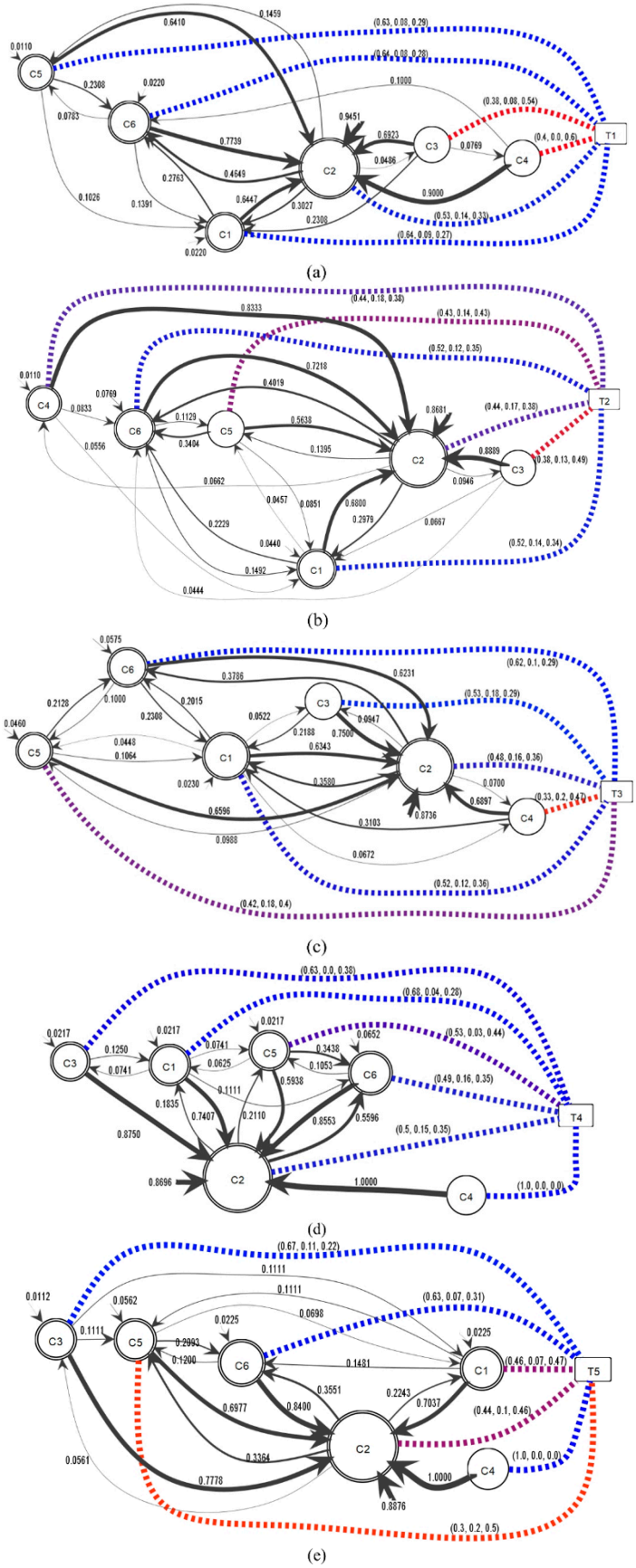

Figure 11. A SentiFlow model for each topic: (**a**) topic t_1; (**b**) topic t_2; (**c**) topic t_3; (**d**) topic t_4; and (**e**) topic t_5.

Figure 11 shows how the topics promote communication between the communities. This answers the first question in this study. For example, Figure 11c presents a description of the information diffusion for topic t_3, with the communication between community c_4 and community c_1 with a probability of 0.3103 and with a response communication probability of 0.0672, which is not observed in the other information diffusion flows. As a response for the second question about how the sentiments are shared from a probabilistic view, the example of information diffusion from community c_2 to community c_4 for topic t_1 shown in Figure 11a is analyzed. Taking the sequence of communities <c_2, c_3, c_4> and the sequence of sentiments <positive, positive, positive>, the probability of the sequence is P (<positive, positive, positive> | Λ) = 0.9451 × 0.53 × 0.0486 × 0.38 × 0.0769 × 0.4 = 2.8455 × 10^{-4}. Moreover, if the communication from community c_2 to community c_6 and the sequence of sentiments <positive, positive> are analyzed, the probability is P (<positive, positive> | Λ) = 0.9451 × 0.53 × 0.4649 × 0.64 = 0.1490 because the only way that c_2 can communicate with c_4 is through c_3, decreasing the probability of the positive sentiment, instead of from c_2 to c_6, where the information diffusion does not need an intermediary community.

5. Conclusions

In this work, an information diffusion process discovery method for SNS was proposed to understand information flow among users better. A SentiFlow model is developed by extending the HMM technique to include process mining by adapting the information from SNS. To understand how the context of user groups is connected with the information flow, different techniques such as an LM for community detection, LDA for natural language preprocessing, and the naïve Bayes classifier for sentiment analysis were used. The proposed method suggested the use of these algorithms, but, in the future, new algorithms can easily be adapted for more accurate and helpful analysis.

The proposed framework has the advantage of allowing users to understand the information flows by displaying the different paths and possible sequences of information delivery obtained from the different users' comments with corresponding probabilities. Analysis of the community of users who plays significant roles in the discovered process shows their sentiment for a related topic.

Three types of information flow diagrams provide the following information. The community information flow describes how the user communities spread their ideas among each other. Moreover, the semantic information flow shown demonstrates how the topics are related with the communities, distinguishing the importance of the topics in each community. Finally, the sentimental information flow presents the potential information to find the focus groups with positive, neutral, or negative opinions and how they influence other user groups according to topic. Additionally, different information diffusion models can be separated and analyzed for each topic.

However, this research still has some limitations. This research focused on understanding the information flow inside a single SNS page, although it can be extended to analyze multiple sites or the whole SNS service. The user profiles of gender, age, and region were not considered in this research, although they may be useful to understand the interactions among users in more detail. In addition, a broader range of human emotions such as anger, joy, and sadness could be used to study the effects of emotions on public opinion. Another limitation is that this research is based on community detection, but the communities may not be stable over time. The study of the reliable community detection can be conducted. In addition, this study focused on the architecture of information diffusion with topic and sentiment, while the analysis methods such as information diffusion process discovery and topic and sentiment analysis were not evaluated. To show the reliability of the analysis result, the detailed methods may be able to be evaluated with evaluation measures such as precision, recall, and F-score.

In future work, a hierarchical model of information flow can be induced to provide different views according to level of abstraction. An integrated approach to capture major interactions among user can be developed without separating the community detection stage and the information flow mining step since the two steps are closely dependent with each other. The dynamics of information flow can also be analyzed to detect the changes of information diffusion in SNS over time.

Author Contributions: B.C. and J.J. conceived and designed the research purposes; B.C. wrote the Proposed Framework and Algorithm Sections; B.C. and J.J. analyzed the data; and B.C. and J.J. wrote the Results and Conclusions Sections.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2013R1A2A2A03014718).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zafarani, R.; Abbasi, M.A.; Liu, H. *Social Media Mining: An Introduction*, 1st ed.; Cambridge University Press: Cambridge, UK, 2014; ISBN 1107018854.
2. Guille, A.; Hacid, H.; Favre, C.; Zighed, D.A. Information diffusion in online social networks: A survey. *Sigmod. Rec.* **2013**, *42*, 17–28. [CrossRef]
3. Grabowicz, P.A.; Ramasco, J.J.; Moro, E.; Pujol, J.M.; Eguiluz, V.M. Social features of online networks: The strength of intermediary ties in online social media. *PLoS ONE* **2012**, *7*. [CrossRef] [PubMed]
4. Tafti, A.; Zotti, R.; Jank, W. Real-time diffusion of information on Twitter and the financial markets. *PLoS ONE* **2016**, *11*, e0159226. [CrossRef] [PubMed]
5. Zhang, X.; Han, D-D.; Yang, R.; Zhang, Z. Users' participation and social influence during information spreading on Twitter. *PLoS ONE* **2017**, *12*. [CrossRef] [PubMed]
6. Jafari, S.; Navidi, H. A Game-Theoretic Approach for Modeling Competitive Diffusion over Social Networks. *Games* **2018**, *9*, 8. [CrossRef]
7. Kim, K.; Jung, J.-Y.; Park, J. Discovery of information diffusion process in social networks. *IEICE Trans. Inf. Syst.* **2012**, *95*, 1539–1542. [CrossRef]
8. Kim, K.; Obregon, J.; Jung, J.-Y. Analyzing information flow and context for Facebook fan pages. *IEICE Trans. Inf. Syst.* **2014**, *97*, 811–814. [CrossRef]
9. Ullah, F.; Lee, S. Social Content Recommendation Based on Spatial-Temporal Aware Diffusion Modeling in Social Networks. *Symmetry* **2016**, *8*, 89. [CrossRef]
10. Kimura, M.; Saito, K.; Nakano, R.; Motoda, H. Extracting influential nodes on a social network for information diffusion. *Data Min. Knowl. Discov.* **2010**, *20*, 70. [CrossRef]
11. Li, D.; Zhang, S.; Sun, X.; Zhou, H.; Li, S.; Li, X. Modeling information diffusion over social networks for temporal dynamic prediction. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 1985–1997. [CrossRef]
12. Kim, M.; Newth, D.; Christen, P. Modeling dynamics of diffusion across heterogeneous social networks: News diffusion in social media. *Entropy* **2013**, *15*, 4215–4242. [CrossRef]
13. Li, M.; Wang, X.; Gao, K.; Zhang, S. A Survey on information diffusion in online social networks: Models and methods. *Information* **2017**, *8*, 118. [CrossRef]
14. Zhao, J.; Dong, L.; Wu, J.; Xu, K. Moodlens: An emoticon-based sentiment analysis system for Chinese tweets. In Proceedings of the 18th ACM SIGKDD, Beijing, China, 12–16 August 2012; pp. 1528–1531.
15. Fan, R.; Zhao, J.; Chen, Y.; Xu, K. Anger is more influential than joy: Sentiment correlation in Weibo. *PLoS ONE* **2014**, *9*, e110184. [CrossRef] [PubMed]
16. Kramer, A.D.; Guillory, J.E.; Hancock, J.T. Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 8788–8790. [CrossRef] [PubMed]
17. Vitale, P.; Guarasci, R.; Iannotta, I.S. Visualizing research topics in Facebook conversations. *Proceedings* **2017**, *1*, 895. [CrossRef]
18. Maynard, D.; Gossen, G.; Funk, A.; Fisichella, M. Should I care about your opinion? Detection of opinion interestingness and dynamics in social media. *Future Internet* **2014**, *6*, 457–481. [CrossRef]
19. Zeng, F.; Zhao, N.; Li, W. Effective social relationship measurement and cluster based routing in mobile opportunistic networks. *Sensors* **2017**, *17*, 1109. [CrossRef] [PubMed]
20. Hernandez-Suarez, A.; Sanchez-Perez, G.; Toscano-Medina, K.; Martinez-Hernandez, V.; Perez-Meana, H.; Olivares-Mercado, J.; Sanchez, V. Social sentiment sensor in Twitter for predicting cyber-attacks using $\ell 1$ regularization. *Sensors* **2018**, *18*, 1380. [CrossRef] [PubMed]
21. Ren, G.; Hong, T. Investigating Online destination images using a topic-based sentiment analysis approach. *Sustainability* **2017**, *9*, 1765. [CrossRef]

22. Bishop, C.M. *Pattern Recognition and Machine Learning*, 1st ed.; Springer: New York, NY, USA, 2006; ISBN 9780387310732.
23. Van der Aalst, W.M.P. *Process Mining: Data Science in Action*, 2nd ed.; Springer: Berlin, Germany, 2016; ISBN 9783662498507.
24. Carrera, B.; Lee, J.; Jung, J.-Y. Discovering information diffusion processes based on hidden Markov models for social network services. In Proceedings of the Asia-Pacific Conference BPM, Busan, Korea, 24–26 June 2015; pp. 170–182.
25. Carrera, B.; Lee, J.; Jung, J.-Y. Discovery of gatekeepers on information diffusion flows using process mining. *Int. J. Ind. Eng.* **2016**, *23*, 253–269.
26. Newman, M. *Networks: An Introduction*, 1st ed.; Oxford University Press: Oxford, UK, 2010; ISBN 9780199206650.
27. Blei, D.M. Probabilistic topic models. *Commun. ACM* **2012**, *55*, 77–84. [CrossRef]
28. Jurka, T. Sentiment: Tools for Sentiment Analysis. R Package Version 02. 2012. Available online: https://github.com/timjurka/sentiment (accessed on 6 March 2018).
29. Liu, B. *Sentiment Analysis and Opinion Mining*; Morgan & Claypool Publishers: San Rafael, CA, USA, 2012.