

Integration of Social Network Analysis and Text Mining for Identification of Influential Opinion Leaders

Chihli Hung¹, Pei-Wen Yeh¹ and Shin-Jye Lee²

Abstract

Opinion leaders are often able to influence others in a specific domain by providing new information, ideas and opinions. Identification of opinion leaders has become one of the most important tasks in the field of word of mouth mining. Most existing research tasks for the identification of opinion leaders are based on quantitative approaches, such as social network analysis, clustering, and statistical methods. However, the content of word of mouth documents is also significant, as opinion leaders often post informative and useful documents with the attributes of novelty, expertise and richness of information. The quality of documents also indicates the possibility of the poster being an opinion leader. Thus, this research proposes two integrated strategies of quantitative and qualitative techniques to evaluate features of opinion leaders. Based on a real-world online discussion forum, the involvement approach is the best of the quantitative approaches. The text mining-based approach and the first integrated approach have the same performance as the involvement approach. The performance of the second proposed integrated approach in identifying opinion leaders achieves the best performance.

JEL classification numbers: M15, O35.

Keywords: Opinion leader, Text mining, Social network analysis, Word of mouth.

¹ Department of Information Management, Chung Yuan Christian University, Taiwan.

² Institute of Management of Technology, National Yang Ming Chiao Tung University, Taiwan.

1. Introduction

Word of mouth (WOM) is a means of communication between an information receiver and a sender, who exchange their experiences of a product or a service (Lakshmi et al., 2025). According to the survey of Keller and Berry (Keller et al., 2025), 83% of people prefer consulting family, friends or experts to traditional advertisements before selecting a new restaurant, and 61% of people seek advice from family, friends or experts before watching movies. People share their experiences on the opinion network. The emergence of the Internet causes WOM to affect the information receivers more quickly, broadly, significantly and without any geographic limitation (Donthu et al., 2021). Potential consumers may save decision-making time and make better decisions by referring to others' opinions on the Internet (Chebil et al., 204). Although making decisions through the opinion network is one of a number of collective decision-making strategies, all kinds of opinion groups have opinion leaders, explicitly or implicitly (Weimann et al., 2007; Zheng et al., 2026). Opinion leaders may bring new information, capture the most representative opinions, and spread their opinions to the masses. Thus, how to identify opinion leaders is crucial and has increasingly attracted the attention of both practitioners and researchers (Kang et al., 2023; Kumar et al., 2023; Song et al., 2007).

Most existing opinion leader identification tasks focus on the number of documents that members initiate, or the number of derived documents and followers for these documents (Agarwal et al., 2008; Miao et al., 2013). Some researchers look for opinion leaders by analyzing the entire opinion network in a specific domain, based on the technique of social network analysis (SNA) (Kim and Han, 2009). These existing approaches in the literature rely mainly on the quantity of documents or network connections, but ignore the significance of quality of documents. For example, a long document is generally more significant than a short one (Agarwal et al., 2008). A document that focuses more on the discussion topic has a higher quality than those documents whose focuses are vague. Thus, documents per se that are posted by members can also be a major indicator for recognizing opinion leaders. This research proposes a hybrid of quantitative and qualitative approaches in order to identify opinion leaders in a real-world bulletin board system. A bulletin board system provides an online forum which represents a type of social network and provides lots of useful WOM information. In this research, novelty, expertise and richness of information from WOM are important features for the identification of opinion leaders. Thus, we propose a text mining-based approach to identify opinion leaders in the bulletin board system. We then integrate the best of the quantitative approaches with the text mining-based approach for the identification of opinion leaders.

The rest of this paper is organized as follows. Section 2 gives an overview of related research work. Section 3 describes the proposed approaches to identify opinion leaders, which include both quantitative and qualitative methods. Section 4 describes the experiment design for this research. The results of the experiments and sensitivity analysis of evaluator consistency are presented in Sections 5 and 6, respectively. Finally, a conclusion and future work are given in Section 7.

2. Related Work

As an opinion leader plays a central role in a social network, a typical network hub plays the same role as an opinion leader. Rosen (2002) uses the acronym *ACTIVE*, to describe the characteristics of typical network hubs i.e. Ahead in adoption, Connected, Travelers, Information-hungry, Vocal, and Exposed to the media more than others. Ahead in adoption means that network hubs may not be the first to adopt new products but they are usually ahead of the rest in the network. Connected means that network hubs play an influential role in a network, such as an information broker among various different groups. Traveler means that network hubs usually love to travel in order to obtain new ideas from other groups. Information-hungry means that network hubs are expected to provide answers to others in their group, so they pursue lots of facts. Vocal means that network hubs like to share their opinions with others and get responses from their audience. Exposed to media means that network hubs open themselves to more communication from mass media, and especially to print media. Thus, a network hub or an opinion leader is not only an influential node but also a novelty early adopter, generator or spreader. An opinion leader has rich expertise in a specific topic and loves to be involved in group activities.

Besides Rosen (2002), who discusses characteristics of opinion leaders in a conceptual way, there are several pieces of research that identify opinion leaders in practical ways. Bamakan et al. (2019) divide methods of opinion leader detection into descriptive methods, statistical methods, diffusion process-based methods, topological-based methods, data mining methods, and learning methods. These methods consist of trust-based relationships, friend-foe relationships, clustering, text mining, sentiment analysis, evolutionary algorithms, social network analysis, ontology-based approaches (Hou, 2022; Hung and Yeh, 2014; Jain et al., 2020; Shang et al., 2021; Zhai et al., 2021; Bhang et al., 2020).

For example, Kim and Han (2009) identify opinion leaders by a two-step process. The first step evaluates degree centrality of members, since members in a social network mutually influence each other. The second step analyzes involvement in activities, which can be done by counting the number of groups that the member belongs to and the number of activities that the member uses. Jiang et al. (2013) propose an improved version of PageRank by using the approaches of sentiment analysis and MapReduce. The PageRank algorithm, proposed by Page et al. (1998), is based on the link structure to evaluate the value of web pages. The core idea of PageRank in the field of opinion leader identification is to identify opinion leaders

due to the number of other members who follow their posts. Kumar et al.(2023) propose a modified spider monkey optimization (Bansal et al., 2014) for opinion leader detection. They use a node2vec (Grover and Leskovec, 2016) method to represent the relationships of nodes in a social network and uses k-means++ to determine both global and local opinion leaders.

Weimann et al. (2007) introduce six approaches to identify opinion leaders in a real society. These are positional, reputational, self-designating, sociometric, observation and key informant approaches. Hudli et al. (2012) observe online behavior of users and define eight attributes that are likely to discriminate opinion leaders from non-opinion leaders. These eight attributes are the time that a user spends online, the frequency with which a user posts messages, the degree to which a user responds to others' posts, the degree to which a user has positive feedback from others, the degree to which a user has negative feedback from others, the degree to which a user refers to others' messages, the average size of messages sent by a user and the degree of involvement of a user.

Song et al. (2007) consider both information novelty and the influential position in the community as criteria for the determination of opinion leaders. Novelty is measured by the similarity between pre-assigned topics and posts. A post is not novel if a document vector is similar to the topic space. Conversely, a non-novel document refers to the documents of many others by a HTML out-link tag. Thus, novelty of documents is proportional to the number of in-links and inversely proportional to the number of out-links (Agarwal et al., 2008). Unlike the work of Song et al. (2007), Li and Du (2011) determine the expertise of authors and readers according to the similarity between their posts and the pre-built term ontology.

Most existing research evaluates opinion leaders using only quantitative approaches but ignoring the significance of published documents per se in a social network (Kumar, Kumar, 2023; Zhou and Zhang, 2023). Zhai et al. (2021) propose a backpropagation neural network for detection of opinion leaders. The sentiment scores of documents from the blogger and the reviewers are used as input features. They use the closeness centrality of the blogger for output result of the neural network. Different from the existing works, we integrate a quantitative approach with a text mining-based approach and compare it with several quantitative approaches. Thus, we identify opinion leaders from both quantitative and qualitative viewpoints.

This research substantially expands upon our preliminary conference paper (Hung & Yeh, 2014). While our earlier work was a comparative study focusing on evaluating single-perspective techniques, i.e., specifically social network analysis versus text mining. This research advances the state-of-the-art by proposing novel integrated models. Rather than relying on a single analytical lens, The integrated models strategically fuse content-based text mining and behavioral involvement. These hybrid approaches address the inherent trade-offs of individual methods, achieving better precision and robustness as confirmed through rigorous inter-rater reliability and sensitivity evaluations.

3. Methodology

This research identifies opinion leaders from both quantitative and qualitative aspects. In the quantitative aspect, we identify the significant features of opinion leaders from the viewpoints of individual and network levels in a specific domain. However, since such techniques usually ignore quality, we extract features of novelty, expertise and richness of information from posts to identify opinion leaders.

3.1 Identifying Opinion Leaders via their Involvement

The first quantitative approach for the identification of opinion leaders is to evaluate their involvement in a social network. Initiating a discussion topic by a post can be considered as a direct involvement in a social network. A derivative post due to this initiated post can be considered as an indirect involvement in a social network. Since involvement is an important characteristic of opinion leaders (Kim and Han, 2009), the number of documents that a member of a social network initiates plus the number of derivative documents by other members is treated as involvement, namely INV, as shown in (1). Members of a social network who have stronger values of involvement are treated as opinion leaders.

$$INV = \text{initiated posts} + \text{derivative posts by other members} \quad (1)$$

3.2 Extracting Significant Network Position Using Social Network Analysis

The INV model evaluates the characteristics of opinion leaders from an individual viewpoint. Social network analysis (SNA) is the analysis of structural relationships between members in a social network. These structural relationships such as strong tie, weak tie, and network centrality, are usually complicated or latent when the number of members is very large. Among these structural relationships, network centrality is often used for the identification of opinion leaders, as it measures the central position for a member in a network (Borgatti, 2006). Freeman (1978) further classify the network centrality into three measures, which are degree centrality, closeness centrality and betweenness centrality.

Degree centrality measures how central a member is based on the number of neighbors surrounding this member. A member with a greater degree centrality indicates a greater representative ability and influence for his or her neighbors. Thus, degree centrality is used for measuring the regional influence of a member within a local area. An opinion leader usually plays a significant role, which affects his or her neighbors. For member i , the equation of degree centrality, DEG , is shown in (2).

$$DEG = \frac{b(n_i)}{g - 1} \quad (2)$$

Closeness centrality (CLO) evaluates the total distance between a member and all other members in a network. A member with a higher closeness centrality indicates that this member has the shortest path to all other members and thus affects them more quickly. An opinion leader usually spreads information more quickly than other members. For member i , the equation of closeness centrality, CLO , is shown in (3).

$$CLO = \frac{g-1}{\sum_{j=1}^g d(n_i, n_j)}, \quad (3)$$

where g indicates the total number of members in a network and $d(n_i, n_j)$ means the shortest distance of members between members i and j .

Betweenness centrality (BET) measures a member that plays the role of a bridge between pairs of other members in a network. A member with high betweenness centrality indicates its bridge position, which is important to the transmission of information in a social network. Hung et al. (2026) treat words as members of a social network and words with high betweenness centrality as members of different semantic communities, thereby achieving word meaning disambiguation (WSD). An opinion leader usually plays such an information-passing role. For node i , the equation of betweenness centrality, BET , is shown in (4).

$$BET = \frac{\sum_{j < k, j \neq i, k \neq i} S_{jk}(n_i) / S_{jk}}{(g-1) \times (g-2)}, \quad (4)$$

where g indicates the total number of members in a network, S_{jk} is the number of the shortest paths that connect members j and k , and $S_{jk}(n_i)$ is the number of the shortest paths that pass member i and connect members j and k .

3.3 Extracting Qualitative Features via Text Mining

Posts per se contain lots of useful information for the identification of opinion leaders, but most existing research ignores their significance (Zheng et al., 2026). Conversely, this research depends on the content of posts to evaluate characteristics of opinion leaders, i.e. novelty, expertise and richness of information, and then integrates the text mining-based approach with traditional quantitative approaches.

3.3.1 Building a Domain Related Lexicon

The main purpose of this initial step is to build a domain related lexicon for the evaluation of novelty, expertise and richness of information, which are important characteristics of opinion leaders. As a Chinese online forum is used, we adopt a normal Chinese text mining process, including Chinese word segmenting, part-of-speech filtering and removal of stop words for the data set of documents. More specifically, the CKIP (Chinese Knowledge and Information Processing) service (<https://ckip.iis.sinica.edu.tw>) is used to segment Chinese documents into proper Chinese words and to search a proper part-of-speech tag for a word from documents.

Only nouns are kept, and a stop list is then used to further remove insignificant words. For unknown words in CKIP, such as English words, we consider them and Chinese nouns as tokens in the text preprocessing stage. Based on these tokens, 85 words are organized into controlled vocabularies. This method is efficient not only to filter irrelevant words but also to capture the main concepts of documents in a specific domain (Gray et al., 2009). Thus, the domain related lexicon vector is represented by a binary vector, i.e. $[I_1, I_2, \dots, I_N]$. As we are interested in the relationships between tokens and members of the online forum, we group documents for each member and build a member by a token matrix as in Table 1. Thus, the member vector $\vec{i}$ is represented as $[O_{i,1}, O_{i,2}, \dots, O_{i,N}]$.

Table 1: Matrix of member by token

	Token 1	Token 2	...	Token N
Member 1	$O_{1,1}$	$O_{1,2}$	...	$O_{1,N}$
Member 2	$O_{2,1}$	$O_{2,2}$	...	$O_{2,N}$
Member 3	$O_{3,1}$	$O_{3,2}$	...	$O_{3,N}$
...	...	...	...	...
Member M	$O_{M,1}$	$O_{M,2}$	...	$O_{M,N}$

O_{ij} denotes the frequency of token j shown in document which is posted by member i .

3.3.2 Novelty

Opinion leaders usually bring new information and start new topics in an opinion network. Finding novelty from the posts by opinion leaders is similar to the task of novelty detection in the field of text mining. Technically speaking, novelty also means the posts which are dissimilar to the normal majority. Approaches of novelty detection can be divided into five categories, which are probabilistic, distance-based, domain-based, reconstruction-based and information theoretic techniques (Pimentel et al., 2014). The distance-based novelty detection is the major approach in text mining. The well-known vector space model transforms a document into a vector and compares similarity of two document vectors with cosine similarity (Salton et al, 1975). Cosine similarity measures the cosine of the angle between two document vectors. Two document vectors are dissimilar when their cosine of the angle is greater. The cosine similarity of two document vectors with the same orientation, i.e. cosine of 0^0 , is 1.

As cosine similarity is more effective on information novelty detection than information theoretic technique, i.e. Kullback Leibler (KL) (Song et al., 2007), we also detect novelty of members by using the cosine similarity function according to their posts. However, a novel post may be an abnormal post as well. We consider novelty as a local situation and evaluate novelty between a post starter and his or her associated followers. Thus, we test the cosine similarity of documents posted by the starters and their followers. In Figure 1, members A, B, C denote document vectors of three post starters. The angles $\theta_A, \theta_B, \theta_C$ denotes the cosine of angle

between documents posted by members A, B, C and his or her associated followers, respectively. Comparing with their associated followers, documents posted by member C is more novel (i.e. more dissimilar) than those posted by members A and B as the cosine of angle between member C and C's followers is greater than others. Novelty of the potential opinion leader i is evaluated as (5). Please note that this is a local function, which tests the difference of documents posted by the starters and their followers. As it is not easy to tell the difference between novel and abnormal posts, we use this local function to limit the opinion leaders who should initiate at least one post.

$$novelty_i = 1 - \frac{1}{F_M} \sum_{x=1, x \neq i}^{F_M} \frac{\sum_{j=1}^N O_{i,j} \times O_{x,j}}{\sqrt{\sum_{j=1}^N (O_{i,j})^2} \times \sqrt{\sum_{j=1}^N (O_{x,j})^2}}, \quad (5)$$

where F_M is the number of followers for member i , N is the number of tokens, and $O_{x,j}$ denotes the frequency of token j shown in a document which is posted by member x .

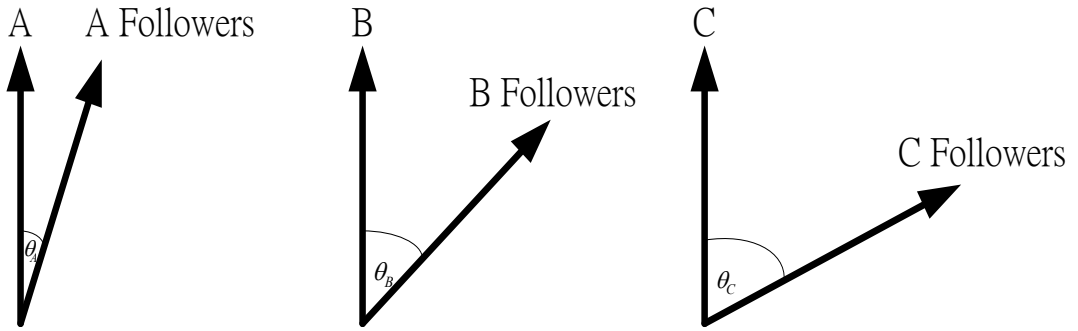


Figure 1: A conceptual example of novelty

3.3.3 Expertise

Opinion leaders usually know more about demographic variables in the topic and present their expertise in their posts (Rosen, 2002). This can be evaluated by comparing their posts with a pre-defined domain vocabulary base (Li and Du, 2011) or the domain related lexicon. Suppose that there are N related tokens in the domain lexicon. For member i , words are collected from his or her posted documents. The term frequency (TF) vector representation approach is used, i.e. $TF_i = [O_{i,1}, O_{i,2}, \dots, O_{i,j}, \dots, O_{i,N}]$, where $O_{i,j}$ indicates the frequency of word j used in the posted documents of user i . In order to avoid the length bias, each member document vector is normalized by his or her maximum frequency of any significant word as in (6). The degree of expertise can be calculated by the cosine similarity function between the member document vector and domain related lexicon vector as shown in (7).

$$f_i = \frac{TF_i}{m_i} = \left[\frac{o_{i,1}}{m_i}, \frac{o_{i,2}}{m_i}, \dots, \frac{o_{i,N}}{m_i} \right], \quad (6)$$

where m_i is the maximum frequency of any significant word in posted documents of user i and N is the total number of tokens in the domain related lexicon.

$$expertise = \frac{\sum_N f_i \times f_v}{\sqrt{\sum_N (f_i)^2} \times \sqrt{\sum_N (f_v)^2}}, \quad (7)$$

where f_i is a vector for user i , f_v is a domain related lexicon vector, and N is the number of significant words in the domain related lexicon.

The degree of expertise can be calculated by the cosine similarity function between the member document vector and domain related lexicon vector. In Figure 2, the angles θ_A , θ_B , θ_C denotes the cosine of angle between documents posted by members A, B, C and the domain related lexicon, respectively. Therefore, degree of expertise for member A is greater than that for members B and C as the cosine of angle between member A and the domain related lexicon is smaller than others.

Domain Related Lexicon

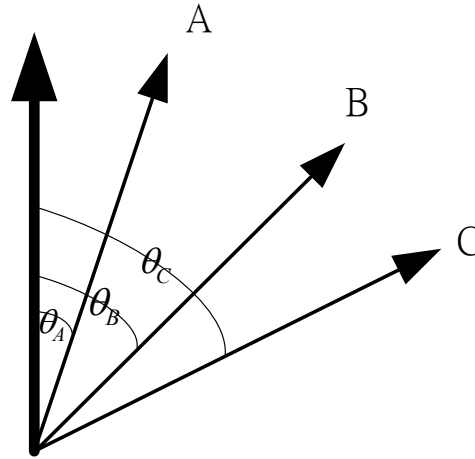


Figure 2: A conceptual example of expertise

3.3.4 Richness of Information

In general, the length of posted documents is positively correlated with the number of comments from readers (Agarwal et al., 2008). In the literature, an influential post is often eloquent (Keller and Berry, 2003). This characteristic can be represented by richness of information. More specifically, we use both textual information and multimedia information to represent the richness of information as $richness_d$ and $richness_m$, respectively, where $richness_d$ is the total number of significant words that the user uses in his or her posts and $richness_m$ is the total number of multimedia objects including anchors that the user posts. Note that each significant word is counted only once and a multimedia object can be found by searching the IMG and A HREF tags in the HTML document.

3.3.5 Text Mining Model

Consequently, we integrate significant features, i.e. novelty, expertise and richness of information including $richness_d$ and $richness_m$, from the content of posted documents, to evaluate the possibility of opinion leaders. As each feature has its own distribution and range, we use a well known min-max normalization to transform each feature into a value between 0 and 1. Thus, the weights for the evaluation of opinion leaders based on the quality of posts become the average of these four features, i.e. novelty, expertise and $richness_d$ and $richness_m$. This approach is defined as the text mining model, namely TXT, which is shown in (8).

$$TXT = \frac{Norm(novelty) + Norm(expertise) + Norm(richness_d) + Norm(richness_m)}{4}, \quad (8)$$

where $Norm(.)$ indicates the min-max normalization function.

3.4 An Integrated Approach

The quantitative approaches identify opinion leaders, which are based mainly on the number of documents that they contribute and their significant positions in a network. The qualitative approaches extract important features of opinion leaders according to documents per se. This research uses four quantitative approaches, i.e. INV, DEG, CLO and BET and one qualitative approach, i.e. TXT. Thus an integrated approach can be formed by combining the qualitative approach with the best quantitative approach. There are two strategies of integration which are used in this research. The first uses a weight setting parameter to adjust significance of the quantitative and qualitative features in order to produce integrated results. This integrated approach, namely INT, is very common, and is shown in (9).

$$INT = \alpha \times Quan + (1 - \alpha) \times Qual, \quad (9)$$

where α is an adjustment whose value is between 0 and 1, $Quan$ indicates weights obtained from a quantitative approach and $Qual$ indicates weights obtained from a qualitative approach.

The second strategy of integration of quantitative and qualitative approaches, namely INT2, is to adopt a two-step method for identifying opinion leaders. Suppose the objective is to identify y opinion leaders. The first step identifies the first x opinion leaders using the qualitative approach, i.e. TXT, and the second step identifies the rest of $y-x$ opinion leaders using the quantitative approach, or vice versa (10). Note that the opinion leaders who have been found by the first approach should be removed from the group. In this research, the approach with the best performance among the quantitative group is chosen. This quantitative approach is then compared with the qualitative approach, i.e. TXT, and the one with the highest performance is used in the first step. When both qualitative and quantitative approaches achieve the same performance, either approach can be used in the first step.

$$INT2 = approach1_x + approach2_{y-x}, \quad (10)$$

where y is the number of opinion leaders and x is an integer between 1 and y . The sample of $approach1_x$ indicates that the first approach is responsible for finding the top x opinion leaders, and $approach2_{y-x}$ indicates that the second approach is responsible for finding the rest of the $y-x$ opinion leaders.

4. Experiment Design

4.1 Data Collection

As there is no available benchmark data set for the purpose of this research, we collect real-world data from the Internet. We take a specific topic, i.e. Nokia, as an example, and crawl data from the Mobile01 (<https://www.mobile01.com/>), which is one of the most popular online discussion forums in Taiwan. Finding opinion leaders on online forums is important in the field of WOM mining, as they are focused on topics that contain a lot of WOM and are easily accessible. We collect 1537 documents, which are initiated by 1064 members. These documents discuss many areas relating to mobile phones and attract 9190 followers, who post 19611 opinions on those initial posts. In this data set, the total number of participants is 9458. Each initial document has 12.76 responses on average.

4.2 Model Comparison

There are several different approaches to identifying opinion leaders in the literature. In this research, we divide them into two groups, i.e. the quantitative approach and the qualitative approach. The number of documents that a member of a social network initiates and the number of derivative documents that are posted by other members are considered as involvement of a member. We consider this method as the first quantitative approach, INV. The second quantitative approach evaluates the possibility of being opinion leaders by using the techniques of social network analysis, including degree centrality, closeness centrality and betweenness centrality, which are called DEG, CLO and BET respectively. This research

presents results of DEG, CLO and BET through the application of pajek (<http://mrvar.fdv.uni-lj.si/pajek>). In terms of the qualitative approach, the research uses the technique of text mining to extract useful information for the identification of opinion leaders from posts. This approach is called TXT for short. We compare four quantitative approaches, i.e. INV, DEG, CLO and BET, and one qualitative approach, i.e. TXT, with human judgment. Finally, to achieve better results, we execute two integrated strategies, i.e. INT and INT2, by combining a best quantitative approach with the qualitative approach.

This research uses real-world data. A user-centered evaluation approach is applied to compare the difference between models. There are 9458 members in this network. We suppose that ten of them have a high possibility of being opinion leaders. In this research, there are four quantitative approaches, which are INV, DEG, CLO and BET, and one qualitative approach, i.e. TXT. The top ten rankings from each model are put in a set of potential opinion leaders. However, many of them are duplicated. We remove duplicates and 25 members are left. The order of these 25 members is shuffled randomly. We request 20 college students who are keen on mobile phone online forums to decide on the possibility of being opinion leaders among these 25 members.

In our questionnaire, information is provided for quantitative consideration, of the number of documents that the potential opinion leaders initiate, and based on these documents the number of derivative documents that are posted by other members. For qualitative consideration, a maximum of three documents from each member are provided randomly to the evaluators. Based on both quantitative and qualitative information, the evaluators evaluate the possibility of being opinion leaders using the Likert 5-point scale. The top 10 rankings are also considered as opinion leaders.

4.3 Evaluation Criterion

The traditional classification criterion, i.e. accuracy, is used in this research as shown in (11). True positive (TP), true negative (TN), false positive (FP), and false negative (FN) are the four different possible outcomes of a single prediction. TP means that a member is predicted to be an opinion leader when this member is an opinion leader. TN means that a member is predicted to be a non-opinion leader when this member is not an opinion leader. Both TP and TN are correct predictions. FP means that a member is incorrectly predicted to be an opinion leader when this member is not an opinion leader. FN means that a member is incorrectly predicted to be a non-opinion leader when this member is an opinion leader.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}, \quad (11)$$

5. Experiment Results

There are 9458 members in our real-world data set. Among these members, we suppose that ten of them are considered as opinion leaders. According to experiment results in Figure 3, TXT and INV outperform other models, followed by BET, CLO and DEG. Among the quantitative approaches, INV performs much better than social network analysis models, i.e. BET, CLO and DEG., even though INV is much simpler than its counterparts from techniques of social network analysis. The major reason may be the characteristic of the network structure, which is very sparse. This is because there are many sub-topics in this online forum, which form several isolated sub-networks. Many members in one sub-network do not provide any comments on other sub-networks. This may detract from the performance of SNA models. Since the best quantitative model is INV, we integrate this model with TXT for the integrated model, INT. We suppose that both quantitative and qualitative parts are equally significant so the value of α is set to 0.5. According to Figure 3, INT only achieves the same performance as INV and TXT.

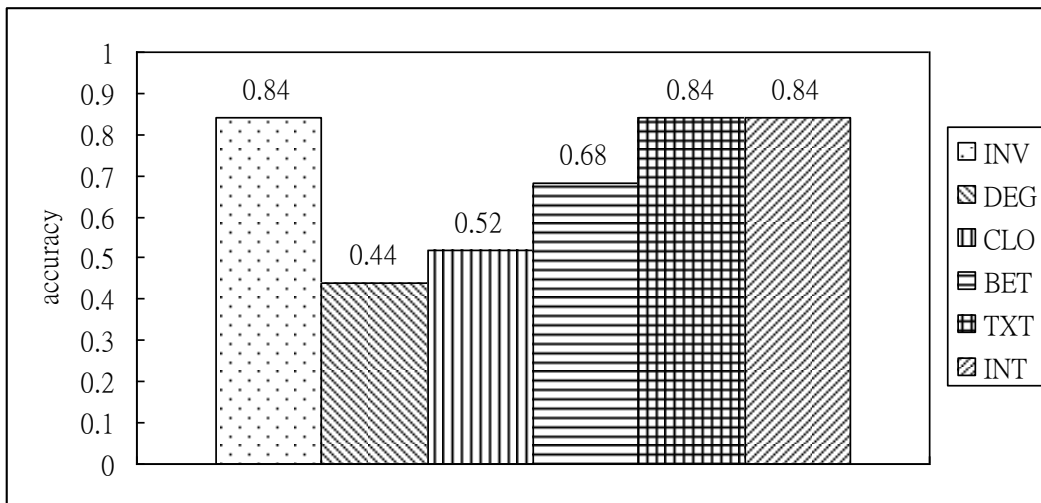


Figure 3: Results of INV, DEG, CLO, BET, TXT and INT models evaluated by accuracy

A possible improvement may be achieved by using different approaches of integration, i.e. INT2. The first approach of INT2 is TXT and the second approach is INV. The value of y is 10 as there are 10 opinion leaders. We set the value of x from 1 to 9 and present the results in Figure 4. A greater value of x indicates that the first approach, i.e. TXT, has a high accuracy in choosing x opinion leaders. The performance of INT2 is the same as the INV, TXT and INT models when x is 1~4 and 8. This proposed integrated model even outperforms the INV, TXT and INT models when x is between 5 and 7.

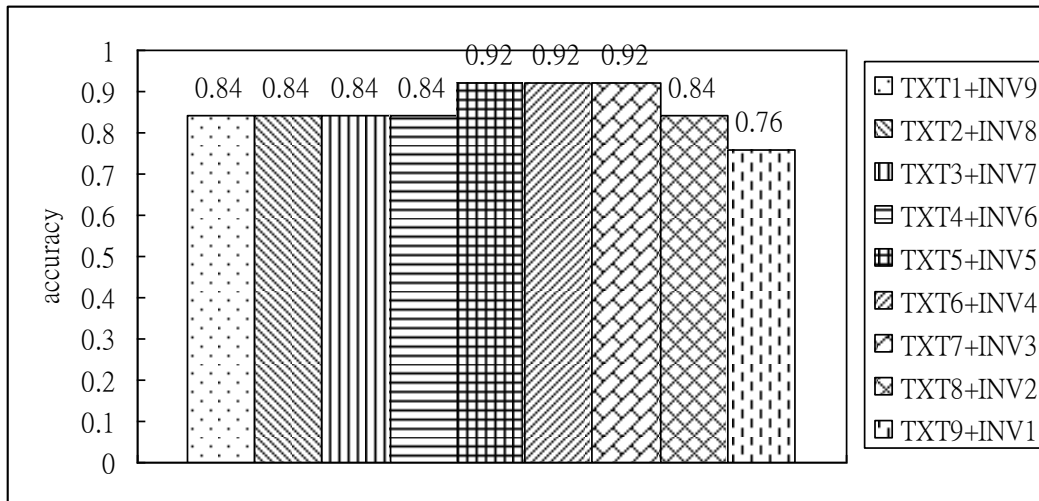


Figure 4: Results of the second integrated model, INT2 which uses TXT first and then followed by INV

As illustrated in Figure 4, the hybrid model INT2 achieves its peak accuracy when the value of x is set between 5 and 7. This specific balance reveals a critical insight into the nature of opinion leaders: true opinion leadership is multidimensional, combining both topical authority and active community engagement. Specifically, text-mining (TXT) and involvement (INV) evaluate influence through distinct criteria. Because the highest-quality candidates naturally concentrate in the upper ranks of both methods, a purely single-method approach suffers from one-sided bias. Setting x to 5~7 effectively forms a balanced hybrid strategy: the initial text-mining phase captures core topical leaders, while the subsequent involvement-based selection complements them with high-engagement community drivers. This 50-50 hybrid synergy demonstrates that comprehensive influencer identification cannot rely solely on what leaders say or how active they are, but rather requires an optimal blend of both dimensions.

6. Sensitivity Analysis of Evaluator Consistency

To establish scientific validity for our human baseline, it is crucial to measure and verify the response consistency among evaluators. Traditionally, inter-rater reliability (IRR) is evaluated using metrics such as Fleiss' Kappa or Krippendorff's Alpha to prove statistically significant agreement among evaluators. However, because our experiment employs a 5-point Likert scale, Krippendorff's Alpha is more appropriate than Fleiss' Kappa. While Fleiss' Kappa treats categorical responses uniformly regardless of distance, Krippendorff's Alpha naturally accommodates interval/ordinal metric weighting, penalizing discrepancies proportionally (e.g., a rating mismatch between 1 and 5 is weighted much heavier than between 4 and 5).

To further test the robustness of our human ground truth and rule out noise from less-consistent raters, we conducted a sensitivity analysis by incrementally filtering out inconsistent evaluators, scaling the pool size from $N = 20$ down to $N = 10$. As summarized in Table 2, Krippendorff's Alpha steadily improves from 0.6162 to 0.6834 as the evaluator pool becomes more refined. Concurrently, we evaluated the criterion of accuracy across these evaluator subsets for all competing models. The results demonstrate that the INT2 model consistently achieves the best performance across all subset configurations. Crucially, the overall conclusions remain robust and consistent regardless of the evaluator subset size, confirming that our findings are highly resilient to individual evaluator variance and reinforcing the superiority of the proposed INT2 framework.

Table 2: Inter-rater reliability (Krippendorff's α) and criterion of accuracy across evaluator subsets

Evaluator Pool Size (N)	Consistent Filter Criterion	Krippendorff's α	INV	TXT	INT	INT2
20	All	0.6162	0.84	0.84	0.84	0.92
18	Top 90%	0.6275	0.80	0.88	0.88	0.88
16	Top 80%	0.6390	0.84	0.84	0.84	0.84
15	Top 75%	0.6412	0.88	0.80	0.80	0.88
14	Top 70%	0.6524	0.88	0.80	0.80	0.88
12	Top 60%	0.6580	0.84	0.84	0.84	0.92
10	Top 50%	0.6834	0.84	0.84	0.84	0.84

7. Conclusion and Future Work

The main purpose of this research is to identify opinion leaders from an online forum. Finding opinion leaders has been of interest to both practitioners and researchers, as opinion leaders play a significant role in diffusion of WOM, which is more credible to consumers than advertisements. Existing models for the identification of opinion leaders mainly focus on quantitative features of opinion leaders, such as the number of posts, influential posts, the central position in the social network, and so on. However, documents per se also provide an important cue. In this research, we extract three important features of opinion leaders, i.e. novelty, expertise and richness of information, from documents. We also compare this text mining-based approach with four quantitative approaches using involvement, degree centrality, closeness centrality and betweenness centrality.

According to our experiments, the involvement approach is the best of the quantitative approaches. This may be because the structure of the network is very sparse and therefore detracts from the performance of those approaches based on social network analysis. The text mining-based approach achieves the same performance as the involvement approach. The main reason may be because the

richness of document information provides a similar function to the qualitative features of opinion leaders. In other words, mining documents provides more significant features of opinion leaders than those pure quantitative approaches based on social network analysis. In order to make an improvement, we propose two integrated strategies of both qualitative and quantitative approaches. The 2-step integrated strategy, which uses the text mining-based approach in the first step, and uses the quantitative approach based on involvement in the second step, achieves the best performance among all of the models.

In terms of possible future work, this research considers the features of novelty, expertise and richness of information to have the same significance in finding opinion leaders. However, this situation may be over-simplified and can be handled by a multi-attribute decision making approach, such as the analytic hierarchy process (AHP) (Saaty and Peniwati, 2008). AHP collects opinions from domain experts and achieves a final weighting agreement for features of opinion leaders. Larger scale experiments including topics, the number of documents and testing, should be done in order to produce more general results. Further research may focus on using alternative integrated approaches and various classification algorithms for identifying opinion leaders. For example, ensemble learning is one possible direction for future investigation.

References

- [1] Lakshmi, L., Ali, L.A.B.M., Devi, K.D.S., Rafiq, M., Shernazarov, I., Othman, N.A. and Khan M.I. (2025). Opinion mining in e-commerce: Evaluating machine learning approaches for sentiment analysis. *Results in Control and Optimization*, 19, no. 100575.
- [2] Keller, E. and Berry, J. (2003). *The Influentials: One American in Ten Tells the Other Nine How to Vote, Where to Eat, and What to Buy*. The Free Press, New York, NY.
- [3] Donthu, N., Kumar, S., Pandey, N., Pandey, N. and Mishra, A. (2021). Mapping the electronic word-of-mouth (eWOM) research: A systematic review and bibliometric analysis. *Journal of Business Research*, 135, pp. 758 - 773.
- [4] Chebil, M., Tobji, M.A.B. and Jallouli, R. (2024). Twitter opinion leaders' detection (TOLD): A comprehensive approach. *Procedia Computer Science*, 246, pp. 1140 - 1149.
- [5] Weimann, G., Tustin, D.H., Vuuren, D. and Joubert, J.P.R. (2007). Looking for opinion leaders: traditional vs. modern measures in traditional societies. *International Journal of Public Opinion Research*, 19(2), pp. 173 - 190.
- [6] Zheng, Y., Wang, Q., Ge, C., Li, Y. and Liu, L. (2026). Reason or emotion? The effects and mechanisms of opinion leader types and mobilization strategies on online collective behavior. *Computers in Human Behavior Reports*, 23, no. 101268.

- [7] Kang, M., Liang, T., Sun, B. and Mao, H.-Y. (2023). Detection of opinion leaders: Static vs. dynamic evaluation in online learning communities. *Heliyon*, 9, no. e14844.
- [8] Kumar, S., Kumar, A., Mallik, A. and S. Dhall. (2023). Opinion leader detection in Asian social networks using modified spider monkey optimization. *ACM Transactions on Asian and Low-Resource Language Information Processing*. 22(5), article 150, pp. 1 - 26.
- [9] Song, X., Chi, Y., Hino, K. and Tseng, B.I. (2007). Identifying opinion leaders in the blogosphere. *Proceedings of 6th ACM International Conference on Information and Knowledge Management*, pp. 971 - 974.
- [10] Agarwal, N., Liu, H., Tang, L. and Yu, P.S. (2008). Identifying the influential bloggers in a community. *Proceedings of International Conference on Web Search and Web Data Mining*, pp. 207 - 217.
- [11] Miao, Q., Zhang, S., Meng, Y. and Yu, H. (2013). Domain-sensitive opinion leader mining from online review communities. *Proceedings of the 22nd International Conference on World Wide Web Companion*, pp. 187 - 188.
- [12] Kim, S. and Han, S. (2009). An analytical way to find influencers on social networks and validate their effects in disseminating social games. *Proceedings of International Conference on Advances in Social Network Analysis and Mining*, pp. 41 - 46.
- [13] Rosen, E. (2002). *The Anatomy of Buzz: How to Create Word of Mouth Marketing*. Doubleday Press, New York, NY.
- [14] Bamakan, S.M.H., Nurgaliev, I. and Qu, Q. (2019). Opinion leader detection: A methodological review. *Expert Systems with Applications*, 115, pp. 200 - 222.
- [15] Hou, L. (2022). Network versus content: The effectiveness in identifying opinion leaders in an online social network with empirical evaluation. *Physica A*, 592, no. 126879.
- [16] Hung, C. and Yeh, P.-W. (2014). Identification of opinion leaders using text mining technique in virtual community. *Proceedings of 1st Symposium on Information Management and Big Data*, pp. 8 - 13.
- [17] Jain, L., Katarya, R. and Sachdeva, S. (2020). Opinion leader detection using whale optimization algorithm in online social network. *Expert Systems with Applications*, 142, no. 113016.
- [18] Shang, Q., Deng, Y. and Cheong, K.H. (2021). Identifying influential nodes in complex networks: Effective distance gravity model. *Information Sciences*, 577, pp. 162 - 179.
- [19] Zhai, Y., Wang, Z., Zeng, H. and Hu, Z. (2021). Social media opinion leader identification based on sentiment analysis. *Proceedings of the 2021 International Conference on Bioinformatics and Intelligent Computing*, pp. 436 - 440.
- [20] Zhang, B., Bai, Y., Zhang, Q., Lian J. and Li, M. (2020). An opinion-leader mining method in social networks with a phased-clustering perspective. *IEEE Access*, 8, pp. 31539 - 31550.

- [21] Jiang, L., Ge, B., Xiao, W. and Gao, M. (2013). BBS opinion leader mining based on an improved PageRank algorithm using MapReduce. *Proceedings of Chinese Automation Congress*, pp. 392 - 396.
- [22] Page, L., Brin, S., Motwani, R. and Winograd, T. (1998). The PageRank citation ranking: Bringing order to the Web. Technical Report, Stanford University.
- [23] Bansal, J.C., Sharma, H., Jadon, S.S. and Clerc, M. (2014). Spider monkey optimization algorithm for numerical optimization. *Memetic Computing*, 6(1), pp. 31 - 47.
- [24] Grover, A. and Leskovec, J. (2016). Node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855 - 864.
- [25] Hudli, S.A., Hudli, A.A. and Hudli, A.V. (2012). Identifying online opinion leaders using k-means clustering. *Proceedings of Intelligent Systems Design and Applications*, pp. 416 - 419.
- [26] Li, F. and Du, T.C. (2011). Who is talking? an ontology-based opinion leader identification framework for word-of-mouth marketing in online social blogs. *Decision Support Systems*, 51, pp. 190 - 197.
- [27] Zhou, X. and Zhang, Z. (2023). Opinion maximization in social networks via leader selection. *Proceedings of ACM Web Conference*, pp. 133 - 142.
- [28] Borgatti, S.P. (2006). Identifying sets of key players in a social network. *Computational, Mathematical and Organizational Theory*, 12, pp. 21 - 34.
- [29] Freeman, L.C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, 1(3), pp. 215 - 239.
- [30] Hung, C., Hung, C.-N. and Chou, H.-M. (2026). Integrating techniques of social network analysis and word embedding for word sense disambiguation. *Kybernetes*, 55(6), pp. 2521 - 2535.
- [31] Gray, A.J.G., Gray, N. and Ounis, I. (2009). Searching and exploring controlled vocabularies. *Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pp. 1 - 5.
- [32] Pimentel, M.A.F., Clifton, D.A., Clifton L. and Tarassnko, L. (2014). A review of novelty detection. *Signal Processing*, 99, pp. 215 - 249.
- [33] Salton, G., Wong, A. and Yang, C.S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), pp. 613 - 620.
- [34] Saaty, T.L. and Peniwati, K. (2008). *Group Decision Making: Drawing Out and Reconciling Differences*, RWS Publications, Pittsburgh, Pennsylvania.