Understanding Effect of Sentiment Content Toward Information Diffusion Pattern in Online Social Networks: A Case Study on TweetScope

Duc Nguyen Trung[1], Tri Tuong Nguyen[1], Jason J. Jung[1(✉)], and Dongjin Choi[2]

[1] Department of Computer Engineering, Yeungnam University, Gyeongsan 712-749, Korea
{duc.nguyentrung,tuongtringuyen,j2jung}@gmail.com
[2] Department of Computer Engineering, Chosun University, Gwangju, Korea

Abstract. Understanding customers' opinion and subjectivity is regarded as an important task in various domains (e.g., marketing). Particularly, with many types of social media (e.g., Twitter and FaceBook), such opinions are propagated to other users and might make a significant influence on them. In this paper, we propose a method for understanding relationship between sentiment content corresponding with its diffusion degree in Online Social Networks. Thereby, a practical system, called *TweetScope*, has been implemented to efficiently collect and analyze all possible tweets from customers.

Keywords: Sentiment analysis · Opinion mining · Online social media · Information diffusion

1 Introduction

It is important for businesses to collect customers' feedbacks about their products and services in direct and more importantly indirect manners [15,16]. Online users have been creating a large amount of information (e.g., personal experiences and opinions) in various forms (e.g., rating [1–3], reviews [5], comments, and articles). Such "personal opinions" among users have be efficiently processed by using various learning methodologies (e.g., decision tree [10], clustering, and so on) [20].

Since many social networking services (SNS) have been emerged, they have enabled the customers to share and exchange their personal opinions. Then, these customers can either make a significant influence on others or get influences from the others [12,13]. For example, if some of friends (or family) have shown any positive (and negative) comments against a certain item (e.g., news and products), one will have a similar feeling regardless of their own personal opinion [6–8,14]. More importantly, the SNS has shown significant power on information diffusion. Once a new piece of information is generated, the information can be propagated to a very large number of other users in a short time.

P.C. Vinh et al. (Eds.): ICCASA 2013, LNICST 128, pp. 349–358, 2014.
DOI: 10.1007/978-3-319-05939-6_34,

The outline of this paper is as follows. Section 2 describes our research issues about the effect of sentiment content toward its diffusion degree from the original source to another social members via their friendship links on Online Social Networks. Sentiment classification method and analysis method for classifying relationship between information diffusion pattern and sentiment content of microtexts posted in Online social networks, are introduced in Sects. 3 and 4. In Sects. 5 and 6, we show the experimental data collection, data preprocessing. The experimental results and evaluations are also discussed. Section 7 draws our conclusion of this work and presents next research directions in the future.

2 Problem Description

A SNS is an open environment where people build their social network or social relations, information from one person can be diffused to another via his/her social links or friendship links. Most social network services are web-based and provide means for user to interact over the Internet, such as Facebook, Google+ and Twitter widely used worldwide. Each SNS has a way of organizing social relation and a mechanism for passing news between a related group of member, Twitter is one of the most popular SNS that enables its users to send and read text-based messages of up to 140 character, know as "tweet" or microtext. Social relations in Twitter is organizing by "*following*" or "*followed*" relationship, a member can follow or be followed by many another members. Followers can read and be notified about tweets, whenever it is posted by who he/she are following, so news can be transmitted among related members. Besides that, Twitter has a powerful and unique news transmission mechanism called "*retweet*" action or RT for short, by which a member can easily copy a tweet from a "following" friend then posts in his/her own timeline. The action has effect to diffuse news from the original author to followers of the author's followers.

In our work, we focused on analysis the relationship between sentiment polarity of tweet content and its diffusion among related Twitter members based *retweet* mechanism. The main research questions are

- Is there any relationship between sentiment content and how information diffuse through social media?
- Is it possible for businesses to employ sentiment opinion to increase effect of the information propagation?

So, there are two main tasks in this work: in the first task, tweets are classified as positive, negative or neutral depend on author's opinion embedded in it; and the second task is to characterize information propagation between Twitter members by measured features, known as information diffusion patterns. Such classification and specification allows us to clarify relationships between distinct sentiment groups of tweets and how it diffused.

For experimenting, we evaluate a case study on a practical system named TweetScope. The application monitors and analyzes data fetched from a text

stream provided by Twitter by filtering tweets on the timeline of certain famous accounts who have a quite enough number of statuses and followers. It's capable of extracting and visualizing the feasible information on marketing [21]. We expect that the visual interface of this system can help decision makers to understand the diffusion patterns on Twitter [9,11].

3 Sentiment Classifier

As description in the above sections, the first task of this work is classification each tweet in to sentiment classes respectively into "positive", "negative" and "neutral". We use common classifier model which has system structure shown as Fig. 1, the model is described detail in the book of Steven Bird et al. [22] with base implemented algorithms packed in a library named "Natural Language Toolkit"[1].

3.1 Naive Bayesian Classifier

The Naive Bayes classifier is often used in text classification due to its speed and simplicity [17–19]. It provides a flexible way for dealing with a number of attributes or classes based on probability theory. The method assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. For given set of classes, it estimated the probability of a class c, given a document d with terms t as

$$P(c|d) = P(c) \prod_{\forall t \in d} P(t|c) \tag{1}$$

Using the rule 1, the Naive Bayes classifier labels a new document as a class c with a decision rule. One common rule is to pick the hypothesis that is most probable, this is know as the maximum a posteriori. The classifier returns a class

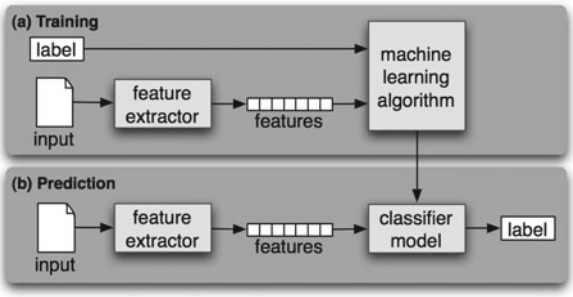

Fig. 1. Common structure of classifier system

[1] Natural Language Toolkit package can be downloaded at http://nltk.org.

level with the highest probability given the document. In our case, the value of class is might be *Positive*, *Negative* or *Neutral*, these classes are assigned a integer value 1, -1 or 0 respectively. In addition if a tweet is granted label *Positive* or *Negative* with a low probability, which is below a threshold ϵ, we considered the tweet is belonged to the neutral class. Therefore, we use following function to classify all tweets in our collected dataset.

$$\text{classify}(d) = \begin{cases} argmax_c\ P(c|d) & \text{if } \exists c\ P(c,d) > \epsilon \\ 0 & \text{Otherwise} \end{cases} \tag{2}$$

4 Diffusion Patterns

4.1 Network Formalization

In order to formalize the information diffusion patterns, we have to define the following notations and show an example. In this paper, we call statuses or news on SNS as microtexts, since they are usually short.

Definition 1 (Microtext). *A microtext twt is a piece of textual information. It is composed of three main features given by*

$$twt = \langle TF, \tau, \Psi \rangle \tag{3}$$

which are (i) term frequencies TF (how many times each term appears in the microtext), (ii) timestamps τ (when the microtext was generated), and (iii) a set of neighbors Ψ (who has been involved in the microtexts).

Given a microtext t, a set of term features $TF(twt, w_k)$ can be extracted by measuring term frequencies in the vector-space model. It can be represented as

$$TF_{twt} = \begin{bmatrix} w_1 & w_2 & \dots & w_{|t|} \\ \frac{count(w_1)}{|t|} & \frac{count(w_2)}{|t|} & \dots & \frac{count(w_{|t|})}{|t|} \end{bmatrix}^T \tag{4}$$

where $w_i \in twt$ is a list of words in twt, and function $count$ returns the number of occurrence of a term w_k. Also, $|twt|$ is the length of twt (i.e., the total number of words).

Definition 2 (Directed Social Network). *A Directed Social Network $\mathcal{S}$ is a network where information is diffused from one to other users via their relationship. It is represented as*

$$\mathcal{S} = \langle \mathcal{U}, \mathcal{N} \rangle \tag{5}$$

where $\mathcal{U}$ is a set of users and $\mathcal{N} \subseteq |\mathcal{U}| \times |\mathcal{U}|$ is a set of relationship betweens the users.

4.2 Diffusion Patterns

In our previous work [4], we have defined a RT network $\mathcal{S}^{twt}_{RT(t)}$ to represent information diffusion network of a specified tweet twt, and also the diffusion pattern of twt by coverage rate ϕ^{twt}.

Definition 3 (Coverage rate). *A coverage rate ϕ can be measured as*

$$\phi^{twt}_{(t)} = \frac{\kappa \times \rho^{twt}_{(t)}}{(1-\kappa) \times \tau^{twt}_{(t)}} \tag{6}$$

where (t) is a certain timestamp for understanding temporal dynamics; $\rho^{twt}_{(t)}$ is a coverage value represent degree of where the target tweet had diffused; $\tau^{twt}_{(t)}$ is a sensitivity known as a response time since the target tweet has been generated, it indicates how quickly users have retweeted [4]. Also, κ is a weighting parameter for emphasizing either coverage (i.e., $0.5 \leq \kappa \leq 1$) or sensitivity (i.e., $0 \leq \kappa < 0.5$).

4.3 Characteristic of Diffusion Patterns

Each Retweet Network of a certain tweet twt has its own a diffusion pattern $g(twt)$ by coverage rate ϕ^{twt} that indicates how many users have diffused twt to others within a unit time. The interesting issue is that some $g(twt)$ can reach its own highest value of ϕ more quick than another; and their maximum value are distributed in various ranges, i.e. In Fig. 2, $g_1(twt_i)$ has a peak later than $g_2(twt_j)$. Besides that, the average value of $g(twt)$ on series of "retweet" times also provide a perspective about the diffusion pattern, so to compare between diffusion patterns, we represent the maximum value of a diffusion pattern $g(twt)$, its time and average value of $g(twt)$ as following

$$\Phi^{twt} = [\phi_{max}, \phi_{avg}, d] \tag{7}$$

where ϕ_{max} is highest value of $g(twt)$ and d is timestamp or retweet position when ϕ_{max} is occurred, ϕ_{avg} is average value of $g(twt)$ over series of "retweet" times.

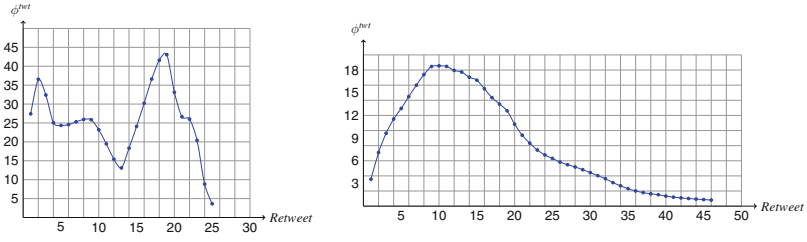

Fig. 2. Two diffusion patterns by (a) $g_1(twt_i)$ of @Windowsphone account, (b) $g_2(twt_j)$ of @SamsumMobile accounts

5 Data Collection

Several famous Twitter accounts which are most popular with high quantity of followers[2], are selected to collect status from 2013, March 1st to 2013, August 1st. Twenty six Twitter accounts are grouped in 3 categories (*Industry*, *Entertainment* and *News*) depend on the accounts' activities. In this section we describe the data, data collection method and data pre-processing step.

5.1 Training Data Set

The performance of classification depends on the subject of data as well as the size of training data set. So, for training the classifier we collect a lot of sentiment microtext from different corpora that are available on previous sentiment analysis studies.

5.2 Data Collection

This work has focused on understanding relationship if any between information diffusion pattern of the tweets and its sentiment content, so there are two subtasks can be done independently: The first one is classifying sentiment class of tweets as positive, negative or neutral based on its content, so does not require any other than pure tweet's text posted on Twitter; the second one is building information diffusion patterns of each tweet. This step require data about relationship between authors of the tweets and who re-tweeted it. Although, these data are available on Twitter and can be fetched easily via Twitter API[3], however there is limits with these functions, so it makes collection task is difficult to obtain data, especially fetching information of retweet action and also friendship relationship. Each Twitter API have a rate limit feature that limit number of the API calling in a small range of time; and number records of returned result is also fixed to each functions. So that, we have no other way to request all necessary data immediately from Twitter rather than building an application to download necessary information from time to time.

5.3 Data Preparation

In this study, each tweet are defined as a list of word which are not enclosed by non-letters to the left and to the right and it need to be considered as not case sensitive for comparing equally. Besides that, tweet usually embedded with some special patterns such as *Short Hyperlink* link to another resource with pattern <*http* || *https*>: // <*Linkcontent*> that describes more detail about

[2] http://twitaholic.com - *Twitterholics* is an online service that scan Twitter a few times a day to determine who is the biggest account.

[3] https://dev.twitter.com/docs provides a detail description about the latest version 1.1 of Twitter API.

content of tweet, but it is not necessary for us analysis even it is good to understand author's opinion; *Mention* is a specified feature of Twitter using pattern @<*username*> that refer to another relevant account however it only help to identity who is mentioned rather than sentiment of the status; and *Hashtag* using pattern #<*hashtag*> that can be used to grouping statuses with similar topic and also giving us some useful information for understand author's opinion e.g., *#romantic*, *#bestphone*, *#incrediblecamera*, etc. Therefore unlike data about information diffusion among users, the content of tweets request to be pre-processed before any analysis task is executed. The data pre-processing is done by the following steps

- *Case-insensitive*: For similar in string comparing method, all tweet is converted to lower case.
- *Hyperlink*: All hyperlink in tweets are replace with a special constant 'LINK'.
- *Mention*: Replace all mentions with a special constant 'USER'.
- *Hashtag*: Because it can contain some useful information for sentiment detection, so we remove symbol '#' from any hashtags to get it original content e.g., '#bestphone' will become 'bestphone'.
- *Numeric*: Numeric string is not prove clearly effect in sentiment expression, so it will remove for the tweets.
- *Special symbol*: Remove all punctuations and other special symbols even if punctuations can help to detect sentiment of the tweet in some cases.
- *Normalization*: The tweets will be rewritten in a form where no space character at either the beginning or end, each words are separated by a single space character.

6 Experimental Results and Discussion

Our *TweetScope* application can collect tweets from Twitter efficiently by using Twitter Stream service; user can access and generate useful data what represent how information diffused from original source to followers on Twitter, Table 1 show a statistic report about our collected dataset.

Using computational result obtained from sentiment classifying task and information diffusion pattern building task, the tweet dataset is classified into

Table 1. Statistics about the data used for the experiment

No		Industry	News	Entertainments
1	Number of twitter accounts	14	5	7
2	Number of tweets	3572	6755	11187
3	Number of feature words	1748	36250	47128
4	Number of stop words	1202	15548	34302
5	Number of hashtags	518	818	981
6	Number of urls	2890	4880	8407
7	Number of mentions	426	849	1703

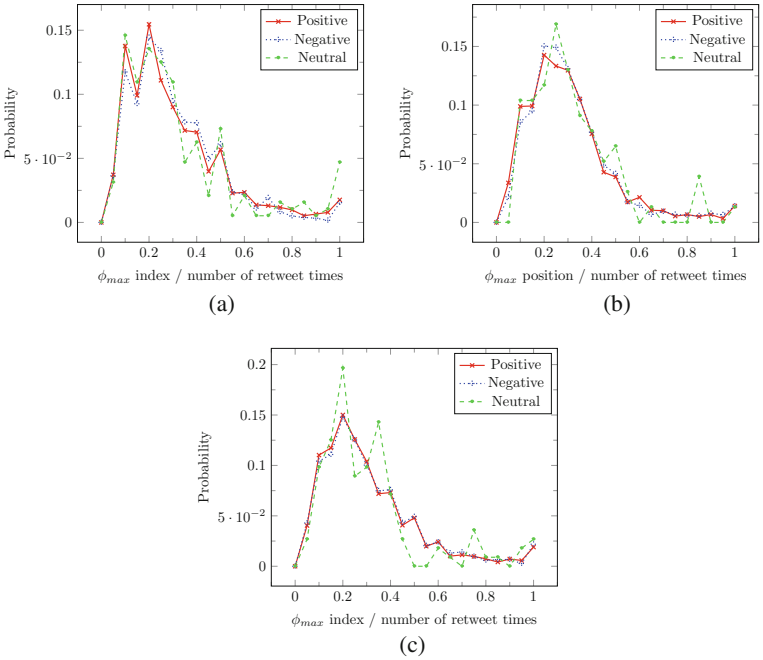

Fig. 3. Probability distribution of ratio between position of ϕ_{max} and number of "retweet" times (a) Industry group, (b) News group, (c) Entertainment group

three group based on its class labels, then we compare difference between characteristic of each polarity groups of tweets. In general we want to compare the data statistically, so ratio between ϕ_{avg}/ϕ_{max} is used as measuring value to show characteristic of diffusion patterns. The ratio value indicates an average rate of information diffused from original source to followers per an unit of time, it is mapped into range [0..1]. By splitting the range [0..1] into a list of small segments, we calculate probability appearance of ϕ_{avg}/ϕ_{max} in each inner range by dividing number of elements in the inner range and number of tweet in each sentiment polarity respectively. These charts indicate that the diffusion patterns of *positive/negative* class have probability of the average rate of information diffusion higher than another of *neutral* class, especially at high value segments. The result is consistent with the assumption that the sentiment content have a certain attraction for related users. However, there are not much difference between chart of classes at some ranges, this can be explained by the influence of noise data, the accuracy of classifier or the novelty of information has its own effect to draw attention of users.

In addition, we consider the time of occurrence of the ϕ_{max} by calculating a factor of retweet index Φ_t^{twt}, where the diffusion pattern get highest peak, and total number of retweet times of the considering tweet. The histogram charts of probability distribution of Φ_t^{twt} / *Number of retweet times* are shown as Fig. 3.

Most of diffusion patterns have a hight probability to reach highest value of coverage rate ϕ_{max} at the one-third of the early stage in series of "retweet" actions. This is quite consistent with the fact that a given news shared on social networks only attracting attention of users in short time around its posting time. The similarity of the graphs shows that shared information on social networks have same chance to spread quickly to related user regardless its sentiment polarity. Once again, may be the novelty of information can play an important role in information diffusion phenomena.

7 Conclusion

In this paper, we developed a method and a practical system to collect and analyze the relationship between sentiment polarity and information diffusion pattern of tweets, which are posted by three group of Twitter users. The experimental result showed a clear potential chance that tweet had sentiment polarity as *positive* or *negative*, have higher probability to diffuse to more users in a given unit of time, even though all kind of tweets are capable to reach highest peak of its diffusion pattern in a short time around posting time regardless its sentiment polarity.

We also recognize that the sentiment polarity and the novelty of news play the same crucial role in attracting attention of users. However, due to the limitations of this work, it will be the target of our next work in the future.

Acknowledgement. This work was supported by the BK21+ Program of the National Research Foundation (NRF) of Korea.

References

1. Brown, J., Broderick, A.J., Lee, N.: Word of mouth communication within online communities: conceptualizing the online social network. J. Interact. Mark. **21**(3), 2–20 (2007)
2. Brunelli, M., Fedrizzi, M.: A fuzzy approach to social network analysis. In: Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining, ASONAM '09, pp. 225–230. IEEE Computer Society, Washington, DC (2009)
3. Dang-Xuan, L., Stieglitz, S.: Impact and diffusion of sentiment in political communication - an empirical analysis of political weblogs. In: Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, pp. 427–430. The AAAI Press, Dublin (2012)
4. Trung, D.N., Jung, J.J., Lee, N., Kim, J.: Thematic analysis by discovering diffusion patterns in social media: an exploratory study with tweetScope. In: Selamat, A., Nguyen, N.T., Haron, H. (eds.) ACIIDS 2013, Part II. LNCS, vol. 7803, pp. 266–274. Springer, Heidelberg (2013)
5. Huffaker, D.: Dimensions of leadership and social influence in online communities. Human Commun. Res. **36**(4), 593–617 (2010)

6. Jung, J.J.: An empirical study on optimizing query transformation on semantic peer-to-peer networks. J. Intel. Fuzzy Syst. **21**(3), 187–195 (2010)
7. Jung, J.J.: Ontology mapping composition for query transformation on distributed environments. Expert Syst. Appl. **37**(12), 8401–8405 (2010)
8. Jung, J.J.: Reusing ontology mappings for query segmentation and routing in semantic peer-to-peer environment. Inf. Sci. **180**(17), 3248–3257 (2010)
9. Jung, J.J.: Service chain-based business alliance formation in service-oriented architecture. Expert Syst. Appl. **38**(3), 2206–2211 (2011)
10. Jung, J.J.: Attribute selection-based recommendation framework for short-head user group: an empirical study by MovieLens and IMDB. Expert Syst. Appl. **39**(4), 4049–4054 (2012)
11. Jung, J.J.: Computational reputation model based on selecting consensus choices: an empirical study on semantic wiki platform. Expert Syst. Appl. **39**(10), 9002–9007 (2012)
12. Jung, J.J.: ContextGrid: a contextual mashup-based collaborative browsing system. Inf. Syst. Front. **14**(4), 953–961 (2012)
13. Jung, J.J.: Discovering community of lingual practice for matching multilingual tags from folksonomies. Comput. J. **55**(3), 337–346 (2012)
14. Jung, J.J.: Evolutionary approach for semantic-based query sampling in large-scale information sources. Inf. Sci. **182**(1), 30–39 (2012)
15. Jung, J.J.: Semantic annotation of cognitive map for knowledge sharing between heterogeneous businesses. Expert Syst. Appl. **39**(5), 1245–1248 (2012)
16. Jung, J.J.: Semantic optimization of query transformation in a large-scale peer-to-peer network. Neurocomputing **88**, 36–41 (2012)
17. Kim, M., Xie, L., Christen, P.: Event diffusion patterns in social media. In: Breslin, J.G., Ellison, N.B., Shanahan, J.G., Tufekci, Z. (eds.) Proceedings of the 6th International Conference on Weblogs and Social Media (ICWSM 2012). The AAAI Press, Dublin (2012)
18. Pham, X.H., Jung, J.J., Hwang, D.: Beating social pulse: understanding information propagation via online social tagging systems. J. Univers. Comput. Sci. **18**(8), 1022–1031 (2012)
19. Salmeron, J.L.: Fuzzy cognitive maps for artificial emotions forecasting. Appl. Soft Comput. **12**(12), 3704–3710 (2012)
20. Sen, S., Lerman, D.: Why are you telling me this? an examination into negative consumer reviews on the web. J. Interact. Mark. **21**(4), 76–94 (2007)
21. Strapparava, C., Mihalcea, R.: Learning to identify emotions in text. In: Proceedings of the 2008 ACM Symposium on Applied Computing, SAC '08, pp. 1556–1560. ACM, New York (2008)
22. Bird, S., Loper, E., Klein, E.: Natural Language Processing with Python. O'Reilly Media Inc., Sebastopol (2009)