From Topics to Opinions::
Modeling Subjectivity for Diffusion Behavior Analysis

Abstract

Information diffusion plays an important role for both researches and applications in current "word of mouth" web. In this paper, we investigate how users' subjectivity influences their information diffusion behavior. Inspired by the psychological research, we define a general subjectivity model by combining both topics and opinions articulated in User-Generated Content (UGC) and propose an efficient framework to establish the subjectivity model. In order to evaluate the impact of subjectivity on information diffusion behavior, a novel subjectivity similarity measurement between two subjectivity models is put forward.The proposed model has been used to predict the retweeting behavior on Twitter, by considering the subjectivity similarities in the attractiveness, sociality and popularity aspects. In the experiments, with a statistical hypothesis test and case study, we demonstrate that a user is more likely to retweet a message in terms of the influence of subjectivity. The evaluation on the practical Twitter dataset shows that our model can improve the performance of retweeting prediction, compared with the state-of-the-art methods.

Introduction

Information diffusion plays an important role in scenes such as marketing and election by triggering and cascading a large number of users. It has drawn considerable attentions from researchers, especially in the area of online social networks. They have built standard models for the general information diffusion, which are useful for simulating the information flow(Goldenberg, Libai, and Muller 2001; Kempe, Kleinberg, and Tardos 2003), or detecting the outbreak of information cascades(Cheng et al. 2014). However, they ignore the intention of users in the process of information diffusion. As information consumer and producer in the web 2.0 era, each user can declare his interests, express opinions, choose information to read and spread on various social media platforms. Whether a message will be spread across the network depends on the "word of mouth" effect, which can elicit the diffusion behavior of users. With the development of Natural Language Processing and data mining techniques, the intention of users can be analyzed by modeling users with user-generated data. In this work, we target at an interesting problem: the mechanism of "word of mouth" effect, i.e. given a new message from a specific user, we intend to predict who will participate in the future diffusion process after receiving this message. As a special scene, we illustrate the problem in a heterogeneous network of Twitter in Figure 1, in which Tony and his friends have tweeted on two topics: cellphone "Iphone" and movie "Frozen". Now Tony posts a new tweet about "Frozen", we want to find out who will disseminate it among all his friends.

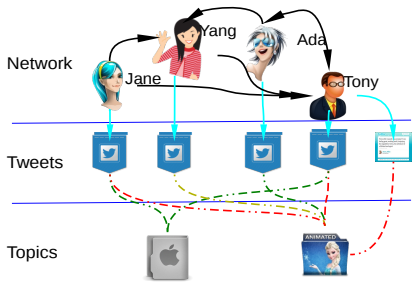

Figure 1: Problem illustration. For opinions of different users, the color "red" stands for positive evaluation, "green" for negative, and "yellow" for neutral.

The diffusion behavior varies in different social networks. In this paper, we investigate our problem under the context of Twitter, which plays an important role in the process of information diffusion on the Internet because the retweeting convention provides an unprecedented mechanism for the spread of information despite the restricted length of a single message. In fact, almost a quarter of the tweets are retweeted from others in Twitter(Yang et al. 2010). Therefore, understanding how retweeting behavior works can help explaining information diffusion in Twitter.

As the participants of information diffusion, users naturally make communication and interaction by expressing opinions and preferences on the topics that interest them. In psychological researches, it has been identified that the subjective initiative nature of human determines that subjectivity will undoubtedly influence his behaviors(Moore and Haggard 2008). According to the theory of Biased Assimilation, people tend to choose and disseminate information according to their own biased opinions(Hyman 2000). Therefore, opinion comprehension is a key aspect of users's inten-

tion in the process of information diffusion. Previous studies have developed a variety of techniques and models to capture the factors that affect retweeting behaviors(Macskassy and Michelson 2011; Feng and Wang 2013). However to the best of our knowledge, few studies have investigated the subjective motivation of a user retweeting a message. For the "word of mouth" effect in Twitter, retweeting is a process that includes receiving, evaluating and deciding whether to share. The crucial part is to evaluate whether a tweet contains information worthy enough to be shared. Therefore modeling the subjective motivation of users will provide an important perspective for retweeting behavior analysis. Intuitively, based on the principle of "like attracts like", a biased user is prone to retweet a message that meets his own tastes. As for the example in Figure 1, the tweets of the users have presented their different opinions on two topics. Tony and Jane were positive on movie "Frozen", while Ada was negative and Yang was neutral. For the new tweet from Tony which is positive on "Frozen", it is easily to understand that Jane is more likely to retweet it because Jane and Tony both like the movie. Therefore how the subjectivity of a user influences his information diffusion behavior is focused on in our work.

For the problem to be investigated, there are two questions arising: how to accurately model the subjectivity of users, and how to effectively measure the worthiness for the users to retweet in terms of subjectivity? Answering the questions is non-trivial. In this paper, we propose a general method to model subjectivity of users, define a novel similarity measurement to calculate the worthiness, and identify factors that influence a user's retweeting behavior from the attractiveness, sociality and popularity aspects.

The rest of the paper is organized as follows: Firstly related works are described; the definition and establishment details of the proposed subjectivity model are given before the subjectivity similarity is defined; Then the factors are specified for the retweeting analysis problem; Further experiments of quantitative evaluation is given; Finally we summarizes the paper and points out future work.

Related Work

Sentiment analysis have tried to mine subjective information from reviews(Liu 2012), and there are also many works pay attention to informal text of social media recently(Jiang et al. 2011; Tan et al. 2011; Guerra, Meira, and Cardie 2014). In our work, instead of focusing on opinion in single text piece, we model subjectivity of a user by integrating all opinions scattering in text pieces to his topics of interest. As another line of works similar to our work, topic-sentiment models can also correlate sentiment with topics, for example, TSM (Mei et al. 2007) model and JST (Lin and He 2009) model. Usually they learn a general word-sentiment distribution to model the sentiment of blogs or reviews, and they represent sentiment as binary polarity. Sentiment expression in informal languages is deemed to be more complicated. Sentiment is often embodied in subtle linguistic characteristics such as: misspellings, abbreviations, emphatic uppercasing, emphatic lengthening and the use of slang and neologisms. Moreover, besides polarity, fine-grained sentiment (strength) is important to distinguish subtle opinions. These lead to much more sparsity in the input and is a special challenge for the word-sentiment distriution. In this paper, we propose to model opinion with a distribution by integrating fine-grained sentiment from a rule-based sentiment analysis method, which can catch subtle sentiment by transforming linguistic characteristics into rules(Thelwall et al. 2010).

A large body of studies has analyzed the characteristics of retweeting behavior(Bian, Yang, and Chua 2014; Luo et al. 2013), examining the factors that lead to increased retweetability(Suh et al. 2010; Comarela et al. 2012) and designing models to estimate the probability of being retweeted(Jenders, Kasneci, and Naumann 2013; Pfitzner, Garas, and Schweitzer 2012). However, all of the above works ignore the subjectivie motivation of users, which is the underlying reason for the retweeting behavior.

Subjectivity Model

With the explosion of social media over the past decade, more and more User-Generated Content (UGC) is available on the Web containing users' opinions. In the Natural Language Processing area, opinion mining techniques(Liu 2012) have been developed to computationally model the subjectivity of users. A variety of aspect-based or topic-sentiment models have been built from UGC by casting opinions as polarity, ratings, or emotions regarding a topic(Lek and Poo 2013; Mei et al. 2007). But their utility is often limited by their representation of opinions. In this paper, we give a more general framework to model subjectivity by combining topics and opinions together with a new representation of opinions. Here we give our definition of subjectivity model under context of Twitter, while we emphasize that our model can be adapted to other context as well.

Definition

Let $G = (V, E)$ denote a social network on Twitter, where V is a set of users, and $E \subset V \times V$ is a set of follow relationships between users. For each user $u \in V$, there is a tweets collection M_u denoting his message history. We assume that there is a topic space T containing all the topics that users in V talk about, and a sentiment space S evaluating their opinions on these topics. For the "subjectivity" of a user $u \in V$, we refer to both topics and opinions articulated in his tweets collection M_u.

Definition 1 (Subjectivity Model) *The subjectivity model of user u, is the combination of topics of interest $\{k\}$ in topic space T and his opinions $\{O_k\}$ on each topic distributed over sentiment space S.*

$$SM_u = \{(k, w_{u,k}, \{d_{u,k,s} | s \in S\}) \, | k \in T\}$$

where:

- *with respect to user u, for each topic $k \in T$, weight $w_{u,k}$ represents the distribution of the user's interests on it, subject to $\sum_{t=1}^{|T|} w_{u,k} = 1$.*
- *opinion of u towards topic k is modeled as a topic-dependent sentiment distribution over sentiment space S, $O_k = \{d_{u,k,s} | s \in S\}$, subject to $\sum_{s=1}^{|S|} d_{u,k,s} = 1$.*

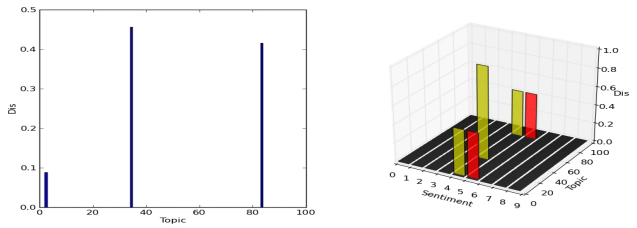

Figure 2: Subjectivity model of user u. The left subgraph denotes interests distribution on topic 2, 32 and 83: $(w_{u,2} = 0.08, w_{u,32} = 0.48, w_{u,83} = 0.44)$. The right subgraph denotes opinions towards topics: $O_2 = (d_{u,2,4} = 0.5, d_{u,2,5} = 0.5)$, $O_{32} = (d_{u,32,4} = 1.0)$, $O_{83} = (d_{u,83,4} = 0.5, d_{u,83,5} = 0.5)$.

The proposed model is general in that the topics of interest and topic-dependent opinions are incorporated into a holistic framework. More importantly, opinions are represented as a probabilistic distribution in a scalable sentiment space which can cover all the sentiment modalities. For example, it could be binary space standing for sentiment polarity, or sequential space for sentiment strength, or discrete space for emotions. Figure 2 is a visualized subjectivity model example in a $[0, 100]$ topic space and a $[0, 8]$ sentiment space.

Establishment of Subjectivity Model

In this section, we present our framework of how to concrete subjectivity model by deriving topics and opinions from the UGC of users $M = \{M_u | u \in V\}$.

Topic Analysis Previous studies have tried to identify topics from tweets by finding key words(Chen et al. 2010), extracting entities(Abel et al. 2011) or linking tweets to external knowledge categories(Macskassy and Michelson 2011). However, works show that topic model is more effective in identifying topics from short and informal social media language(Hong and Davison 2010). In this paper, state-of-the-art Latent Dirichlet Allocation (LDA)(Blei, Ng, and Jordan 2003) is employed for unsupervised topic discovery and for topic assignment of future tweets. LDA can be used to find a set of K latent topics from a document corpus, and then to represent each document D with a distribution θ_D of the latent topics. For each word w_i in D, a topic z is first sampled from the document topic distribution θ_D, then w_i is sampled according to word distribution ϕ_z of topic z.

As the first step, we adopt the user-level LDA model to build a global Topic Model (TM), which takes all tweets of a user M_u as one document of LDA(Hong and Davison 2010). The TM model will be used throughout our framework. Usually a tweet concentrates on a single topic within its short length, therefore we assign a tweet t to a topic that maximizes the probability of generating t:

$$z_t = \arg\max_k \prod_{w \in t} P(w|\phi_k) \tag{1}$$

We can get a weight distribution on each topic k of user u by normalizing all tweets that talk about topic k:

$$w_{u,k} = \frac{|\{t : t \in M_u \wedge z_t = k\}|}{|M_u|} \tag{2}$$

Opinion Analysis Considering the example in Figure 1, both Tony and Jane held an overall positive opinion on the movie "Frozen", but maybe they liked the movie for different reasons. Jane maybe mainly liked the romantic story of this movie but was a little disappointed at its animation picture, while Tony liked this movie perhaps because he was mostly convinced by its animation technology although he disliked the prince and princess genre. Previous works usually represent opinion with a simple binary polarity, without differentiating opinions on different aspects, thus they may not satisfy the information discovery needs of different users. Therefore, it is better to describe opinion for a topic as a probability distribution over the sentiment space. Furthermore, a more fine-grained sentiment space is preferred if we want to distinguish subjectivities of users more precisely.

Researches on sentiment analysis of social media have provided many effective state-of-the-art techniques and tools(Thelwall et al. 2010; Hu et al. 2013), with which sentiment of a tweet t can be identified as s_t. The opinion distribution O_k toward a topic k can be integrated as:

$$\begin{aligned} O_k &= \{d_{u,k,s} | s \in S\} \\ &= \{\frac{|t : t \in M_u \wedge z_t = k \wedge s_t = s|}{|M_u|} | s \in S\} \end{aligned} \tag{3}$$

Subjectivity Similarity

With the subjectivity model established, a subjectivity similarity measurement needs to be calculated to analyze various subjective decision-making processes such as retweeting behavior. Firstly we should define the opinion similarity on a common topic.

Opinion Similarity Opinion in the subjectivity model is treated as a distribution over sentiment space with each dimension of the distribution representing the proportion of the corresponding sentiment value. In fact, values of the sentiment space are not independent. They are sequential in magnitude and quantified to measure the strength of sentiment. Therefore, normal distribution similarity measurements such as KL-divergence and cosine similarity are not suitable for such kind of opinion distribution. As illustrated in Table 1, in a $S = [0, 8]$ integer sentiment space, opinion O_k^1 is most negative (100% of value 0), opinion O_k^2 (50% of value 6 and 50% of value 7) is positive, and O_k^3 (100% of value 8) is most positive. If the cosine similarity measurement is adopted, all similarities among them are 0. In fact O_k^2 is more similar to O_k^3 than O_k^1 because they both are positive and their strength distance is much less than O_k^1. Therefore, opinion similarity can't be calculated simply as normal probabilistic distributions, or just as strength distance. To accurately catch opinion similarity, we propose a novel method by combining strength distance and distribution similarity. The opinion similarity between two opinions O_k^u, O_k^von the same topic k can be calculated as:

$$Sim(O_k^u, O_k^v) = \frac{|S| - |\sum_{i=0}^{|S|} d_i^u v_i - \sum_{i=0}^{|S|} d_i^v v_i|}{|S|} \tag{4}$$

where d_i denotes the i^{th} dimension of opinion distribution, and v_i denotes corresponding sentiment value. The

Table 1: Illustration of opinion similarity

	0	1	2	3	4	5	6	7	8
O_k^1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
O_k^2	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.5	0.0
O_k^3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

similarities of opinions in Table 1 calculated with Equation 4 are $Sim(O_k^1, O_k^3) = 0$, $Sim(O_k^2, O_k^3) = 6/8$ and $Sim(O_k^1, O_k^2) = 2/8$, which are consistent with our intuitive understanding.

Subjectivity Similarity As the subjectivity model indicates, a user's topics of interest is represented as a weight distribution over topic space T. Therefore, the subjectivity similarity between two subjecctivity models SM_u and SM_v should be integrated by combining the topic weight and the opinion similarity on each common topic:

$$Sim(SM_u, SM_v) = \sum_{k=1}^{|T_{u,v}|} \theta_u(k) Sim(O_k^u, O_k^v) \quad (5)$$

Where $T_{u,v}$ denotes the common topics between two users, which can be regarded as the intersection of their topics of interest; $\theta_u(k)$ denotes the weight of topic k of user u.

Note that, when we measure how similar user u is to user v, we use the topic weight of user u, thus the subjectivity similarity is asymmetric. The intuition lies in that subjectivity of a user is a personal inner taste, and it is also one-way judgement about how like-minded a friend is. Therefore, for the measurement of subjectivity similarity, $Sim(SM_u, SM_v) \neq Sim(SM_v, SM_u)$.

Retweeting Analysis

Whether a user retweets a message may be affected by various factors. From the point of a user, three situations usually make him rewteet: 1) the content of the tweet is attractive for the user, and his retweeting behavior is in accordance with subjective evaluation; 2) the tweet is posted by the user's close friend, and his retweeting behavior is due to social needs; and 3) the content is popular, and his retweeting behavior is a result of conformity needs(Cialdini and Goldstein 2004). These situations exhibit different types of reasons of rewteeting behavior, and we quantify them with three subjectivity similarities for the motivation analysis in our work. For a target tweet t, let F denote the followers who receive t by following its author u_a. For each follower $f \in F$, we can define a quadruple $< f, u_a, t, r_f >$, where r_f is a binary label indicating if t is retweeted by f, which needs to be predicted.

Attractiveness

A user is likely to repost a tweet if the user finds the content is attractive according to his subjective judgement. We can measure such attractiveness quantitatively by calculating the subjectivity similarity between the tweet t and user f. For a tweet t, its topic z_t can be identified with Equation 1, and let s_t be its sentiment. The content of t can also be modeled using subjectivity model definition with a single topic of interest and a 100% opinion distribution on sentiment value s_t. Thus the attractiveness of tweet t to user f can be measured with subjectivity similarity using Equation 5, which is marked as $Sim(f, t)$.

Sociality

In this case, the retweeting behavior is based on the needs of social interaction. That is, the behavior is triggered because the tweet is sent by a like-minded friend, instead of the information it contains. We can measure how like-minded the user f and his friend u_a are with their subjectivity similarity $Sim(f, u_a)$. However, different kinds of friends may have a different influence on the user f. For example, f may follow many friends, but only frequently interacts with a few. Furthermore, not all tweets of a friend may be of interest to f. For example, in Figure 1, Jane may be interested in the tweets from Tony about movie, but not interested in his tweets about cellphone. We therefore assign a weight to $Sim(f, u_a)$ to reflect the influence of different kinds of friend, which is composed of four factors:

Expert Factor $w_E(u_a)$: It represents the relative expertise of the author u_a among his followers including f. The expert user imposes more influence on others. We simply calculate it as the ratio of user u_a's tweets count over all tweets of u_a and his followers by $w_E(u_a) = |M_{u_a}|/|\{M_u | u \in u_a \cup F\}|$.

Leadership Factor $w_L(u_a)$: In our work, the leadership of a user u_a is determined by his followers. The leadership weight is calculated by $w_L(u_a) = \log(|F|)/\log(\max)$, where $\max$ is the maximum popularity of a user in Twitter[1].

Similarity Factor $w_S(u_a, f)$: The similarity of interests between u_a and f is measured as the inverse KL-divergence between their topic weight distribution in their subjectivity model: $w_S(u_a, f) = 1/KL(\theta_{u_a}, \theta_f)$.

Interaction Factor $w_I(u_a, f)$: All the interactions $Interation_{u_a,f}$ between u_a and f are analyzed, which include the conversations between them, mentions of each other, and retweets from each other. The factor weight is calculated by normalizing $Interation_{u_a,f}$ with all tweets of u_a and f: $w_I(u_a, f) = |Interation_{u_a,f}|/|\{M_{u_a}, M_f\}|$.

Above all, the influence weight is the combination of four factors:

$$\begin{aligned} w_{u_a,f} =& \lambda_1 * w_E(u_a) + \lambda_2 * w_L(u_a) + \\ & \lambda_3 * w_S(u_a, f) + \lambda_4 * w_I(u_a, f). \end{aligned} \quad (6)$$

where λ_i is an optional weight vector to enable different influence of the factors, subject to $\sum_{i=1}^{4} \lambda_i = 1$. We set them uniformly as 0.25 in our work.

Popularity

If a tweet is popular(novel or epidemic), it will be very probable to be retweeted. In this situation, the tweet t is often

[1] http://twittercounter.com/pages/100

inconsistent with the interests and opinions of its author u_a. Thus the similarity between t and u_a in terms of subjectivity is relatively low, which is marked as $Sim(u_a, t)$. The retweeting behavior is highly related to the popularity of t rather than the content or the friend who post it. We assign a popularity weight to $Sim(u_a, t)$, which is the proportion of user f's followees who have retweeted the tweet t.

From the point of motivation, a user might retweet a message if its content is approximate to his subjectivity, its author is a like-minded friend or it is popular among his friends.

Experiments

Dataset and Settings

Our method has been evaluated on the Twitter dataset from (Luo et al. 2013), in which 500 target English tweets were monitored to find who would retweet it in the future. Each target tweet was set as starting point to collect recent tweets for its author and followers. Overall, there are 45,531 users who have posted at least 6,277,736 tweets. 5214 users have retweeted at least one target tweet during the monitored period. To avoid the bias introduced by dataset imbalance, an evaluation dataset is constructed by taking 5,214 retweeters as positive instances, and randomly sampling 5,214 non-retweeters as negative instances.

For the topic model, we use Gensim(Řehůřek and Sojka 2010), which adopts an efficient batch-based online inference algorithm. All parameters are set as defaults and the number of topic traverses from 50 to 200. For sentiment analysis, we just make use of an off-the-shelf work, i.e. SentiStrength(Thelwall et al. 2010). In order to catch the sentiment of tweets, we use the sentiment lexicon created based on AFINN by Nielsen(Mohammad, Kiritchenko, and Zhu 2013). The sentiment space is formed by mapping the positive and negative sentiment values to range $[0, 8]$.

Correlation Test

First of all we assess the correlation between subjectivity similarity and retweeting behavior with a statistical hypothesis test Analysis of Variance (ANOVA)(Fisher et al. 1970), which tests the *null hypothesis* that the retweeters and non-retweeters have the same subjectivity similarity means. The results are listed in Table 2. The bold-faced entries mean that the *p-value* is lower than significance level. Note that for the topic numbers of 100 and 150, all similarities yield *p-values* below significance level with *F-ratio* above 1.00. This suggests that the subjectivity similarities could be useful features for modeling retweeting behavior. For the rest experiments, we set the topic number as 100 for LDA.

Table 2: ANOVA results for subjectivity similarities. If the difference is due to chance, *F-ratio*=1.00, otherwise *F-ratio* >1.00 (*p-value* <0.01).

Similarity		$Sim(f,t)$	$Sim(f,u_a)$	$Sim(u_a,t)$
50	*F*	**12.182**	2.212	4.236
	p	$\mathbf{4.44e^{-06}}$	0.140	0.272
100	*F*	**43.892**	**31.145**	**28.466**
	p	$\mathbf{8.65e^{-11}}$	$\mathbf{3.55e^{-08}}$	$\mathbf{1.32e^{-09}}$
150	*F*	**22.356**	**12.240**	**14.664**
	p	$\mathbf{2.43e^{-08}}$	$\mathbf{6.25e^{-06}}$	$\mathbf{8.46e^{-07}}$
200	*F*	**31.675**	**20.616**	6.145
	p	$\mathbf{4.22e^{-06}}$	$\mathbf{2.92e^{-05}}$	0.26

Case Study

In this section, we give an vivid example to illustrate the subjectivity model and its ability in explaining the retweeting behavior. The subjectivity models of one of the 500 target tweets, its author, and two followers (one retweeter, the other non-retweeter) are shown in Figure 3. The right part of each sub-figure illustrates topic distribution and the left part illustrates opinions on each topic.

It is obvious that the tweet is about the 14^{th} topic, and the opinion is neutral. The author concentrates on the 14^{th} topic, and his opinion is mainly neutral. As for two followers, the retweeter has tweeted on two topics (the 14^{th} and 52^{nd} topic) uniformly and his opinion on the 14^{th} topic is mainly neutral. While the non-retweeter has also talked about two topics (14^{th} and 56^{th} topic), but he is mainly interested in the 14^{th} topic and his opinion is positive.

Table 3 shows the three subjectivity similarities for both retweeter and non-retweeter. It is clear that except for the similarity between the tweet and its author, the other two subjectivity similarities of the retweeter are much larger than the non-retweeter. They have common interest (the 14^{th} topic), and furthermore the non-retweeter is more similar with the tweet and its author than the retweeter in terms of topics. But their different opinions towards the topic elicit their different behaviors, which verifies our model can help better understanding the retweeting behavior not only from topics but also opinions.

Table 3: Illustration of example subjectivity similarities

Similarity	$Sim(f,t)$	$Sim(f,u_a)$	$Sim(u_a,t)$
Retweeter	0.854	0.967	0.886
Non-retweeter	0.805	0.919	0.886

Performance Evaluation

We carried out the retweeting prediction experiments in three stages. Firstly we compared our model against other topic-based models including TF-IDF model (modeling user interests using bag-of-words), entity-based model (using entities extracted from the UGC) and hashtag-based model (using hashtags used in the UGC)(Abel et al. 2011). Secondly, our model was compared with two generative topic-sentiment models (TSM model(Mei et al. 2007) and JST model(Lin and He 2009)). TSM and JST can also model topic and topic related sentiment simultaneously. We also use Equation 5 to calculate three subjectivity similarities for bith TSM and JST as our method in section , and combine them together in the predction.

The subjectivity model has been proposed to catch the subjective motivation of users based on UGC, whereas other important factors associated with retweeting behavior are not considered, such as network topology and meta-data

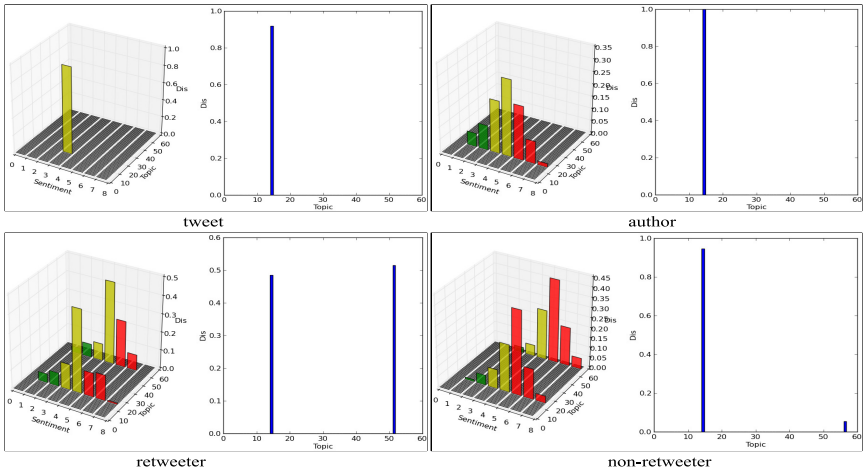

Figure 3: An illustration of subjectivity models of a tweet, author and two followers.

of users. Therefore, our model is also compared with the method of Luo *et al.* (2013)(marked as "LUO"), in which different factors that might affect rewteeting behaviors are considered. They only use bag-of-words to model user interests, so we also carried out combining experiments to demonstrate that the performance of prediction can be improved by replacing their bag-of-words model with our model (marks with "LUO+" prefix).

Table 4: Accuracy performance. A significant improvement over baseline with $*$ and LUO's model with ‡ ($p < 0.05$).

Feature	Accuracy(%)	Feature	Accuracy(%)
baseline	60.85		
TF-IDF	62.85 $*$	LUO	71.76 $*$
entity	68.76 $*$	LUO+entity	72.15 $*$
hashtag	59.12	LUO+hashtag	68.44 $*$
TSM	67.44 $*$	LUO+TSM	68.23 $*$
JST	68.13 $*$	LUO+JST	70.53 $*$
$Sim(f,t)$	73.88 $*$ ‡	LUO+$Sim(f,t)$	74.04 $*$ ‡
$Sim(f,u_a)$	70.04 $*$	LUO+$Sim(f,u_a)$	70.27 $*$
$Sim(u_a,t)$	69.64 $*$	LUO+$Sim(u_a,t)$	71.86 $*$
sim_{all}	**75.64** $*$ ‡	LUO+sim_{all}	**78.15** $*$ ‡

The logistic regression classifier is used for training and testing in a 5-fold cross-validation manner. We set a baseline, which simply predicts users who have retweeted the author previously as the retweeters of target tweet. All results are presented in Table 4 in terms of accuracy.

Firstly, all models except the hashtag-based model outperform the baseline (60.85%) significantly. While for hashtag-based model, the accuracy is only 59.12%, the reason lies in the sparcity of hashtag in tweets.

Secondly, $Sim(f,t)$ and sim_{all} outperform "LUO" (71.76%) significantly. The best performance is achieved by the sim_{all} (75.64%), for which we add three similarities to the classifier to test the impact of their combination. The performance of TF-IDF model (62.85%) is a little better than baseline. The entity-based model (68.76%) is very close to $Sim(f,u_a)$ (70.04%) and $Sim(u_a,t)$ (69.64%), and the difference is not significant.

Thirdly, the performance of two topic-sentiment models (TSM: 67.44%, JST: 68.13%) is not as good as our models. The reason lies in that they use a binary sentiment representation (positive or negative), which can not differentiate opinions elaborately. Our model can capture more subtle and fine-grain sentiment, which could distinguish different subective motivation of retweeting behavior.

Finally, in the combining evaluation, $Sim(f,t)$ gives a significant improvement (LUO+$Sim(f,t)$, 2.12% improvement) over "LUO", but other two similarities and the entity-based model can not improve performance significantly. The performance is even degraded after combining with the hashtag-based model and two topic-sentiment models. But noticing that, the most significant improvement(LUO+sim_{all}, 6.39% improvement) is achieved by combining with all three similarities.

Above all, the results show that our model can better help predicting retweeting behavior and can be regarded as a useful way to analyze the retweeting behaviors of users.

Conclusion

Motivated by the psychological research, this paper postulates that the diffusion behaviors of social media users are affected by their subjectivity. Therefore, a general subjectivity model has been proposed and an efficient framework has been designed to establish the subjectivity model. Also a novel method is proposed to measure the subjectivity similarity. The subjectivity model has been applied to the retweeting analysis considering the attractiveness, sociality and popularity factors, which are quantified with subjectivity similarities. Experiment results demonstrate the effectiveness of the proposed model in the retweeting analysis problem and show that the model is able to reach better understanding of retweeting behavior. In the future, we will apply our model to other social network analysis task such as link prediction and recommendation.

References

Abel, F.; Gao, Q.; Houben, G.-J.; and Tao, K. 2011. Analyzing user modeling on twitter for personalized news recommendations. In *UMAP*. Springer. 1–12.

Bian, J.; Yang, Y.; and Chua, T.-S. 2014. Predicting trending messages and diffusion participants in microblogging network. In *Proc. of SIGIR '14*, SIGIR '14, 537–546. New York, NY, USA: ACM.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Chen, J.; Nairn, R.; Nelson, L.; Bernstein, M.; and Chi, E. 2010. Short and tweet: experiments on recommending content from information streams. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, 1185–1194. ACM.

Cheng, J.; Adamic, L.; Dow, P. A.; Kleinberg, J. M.; and Leskovec, J. 2014. Can cascades be predicted? In *Proc. of the 23rd WWW*, 925–936. WWW.

Cialdini, R. B., and Goldstein, N. J. 2004. Social influence: Compliance and conformity. *Annu. Rev. Psychol.* 55:591–621.

Comarela, G.; Crovella, M.; Almeida, V.; and Benevenuto, F. 2012. Understanding factors that affect response rates in twitter. In *Proc. of the 23rd ACM conference on Hypertext and social media*, 123–132. ACM.

Feng, W., and Wang, J. 2013. Retweet or not?: personalized tweet re-ranking. In *Proc. of the 6th WSDM*, 577–586. ACM.

Fisher, S. R. A.; Genetiker, S.; Fisher, R. A.; Genetician, S.; Britain, G.; Fisher, R. A.; and Généticien, S. 1970. *Statistical methods for research workers*, volume 14. Oliver and Boyd Edinburgh.

Goldenberg, J.; Libai, B.; and Muller, E. 2001. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters* 12(3):211–223.

Guerra, P. C.; Meira, Jr., W.; and Cardie, C. 2014. Sentiment analysis on evolving social streams: How self-report imbalances can help. In *Proc. of the 7th WSDM*, WSDM '14, 443–452. New York, NY, USA: ACM.

Hong, L., and Davison, B. D. 2010. Empirical study of topic modeling in twitter. In *Proc. of the First Workshop on Social Media Analytics*, 80–88. ACM.

Hu, X.; Tang, J.; Gao, H.; and Liu, H. 2013. Unsupervised sentiment analysis with emotional signals. In *Proc. of the 22nd WWW*, 607–618. WWW.

Hyman, J. 2000. Three Fallacies about Action. *Behavioral and Brain Sciences* 23:665–666.

Jenders, M.; Kasneci, G.; and Naumann, F. 2013. Analyzing and predicting viral tweets. In *Proc. of the 22nd WWW*, 657–664. WWW.

Jiang, L.; Yu, M.; Zhou, M.; Liu, X.; and Zhao, T. 2011. Target-dependent twitter sentiment classification. In *Proc. of the 49th ACL*, 151–160. ACL.

Kempe, D.; Kleinberg, J.; and Tardos, É. 2003. Maximizing the spread of influence through a social network. In *Proc. of the ninth ACM SIGKDD*, 137–146. ACM.

Lek, H. H., and Poo, D. C. 2013. Aspect-based twitter sentiment classification. In *Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference*, 366–373. IEEE.

Lin, C., and He, Y. 2009. Joint sentiment/topic model for sentiment analysis. In *Proc. of the 18th ACM CIKM*, 375–384. ACM.

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies* 5(1):1–167.

Luo, Z.; Osborne, M.; Tang, J.; and Wang, T. 2013. Who will retweet me?: finding retweeters in twitter. In *Proc. of the 36th international ACM SIGIR*, SIGIR '13, 869–872. New York, NY, USA: ACM.

Macskassy, S. A., and Michelson, M. 2011. Why do people retweet? anti-homophily wins the day! In *ICWSM*.

Mei, Q.; Ling, X.; Wondra, M.; Su, H.; and Zhai, C. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proc. of the 16th WWW*, 171–180. ACM.

Mohammad, S. M.; Kiritchenko, S.; and Zhu, X. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proc. of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*.

Moore, J., and Haggard, P. 2008. Awareness of action: Inference and prediction. *Consciousness and cognition* 17(1):136–144.

Pfitzner, R.; Garas, A.; and Schweitzer, F. 2012. Emotional divergence influences information spreading in twitter. In *ICWSM*.

Řehůřek, R., and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proc. of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

Suh, B.; Hong, L.; Pirolli, P.; and Chi, E. H. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *2010 IEEE Second International Conference on Social Computing*, 177–184. IEEE.

Tan, C.; Lee, L.; Tang, J.; Jiang, L.; Zhou, M.; and Li, P. 2011. User-level sentiment analysis incorporating social networks. In *Proc. of the 17th ACM SIGKDD*, 1397–1405. ACM.

Thelwall, M.; Buckley, K.; Paltoglou, G.; Cai, D.; and Kappas, A. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* 61(12):2544–2558.

Yang, Z.; Guo, J.; Cai, K.; Tang, J.; Li, J.; Zhang, L.; and Su, Z. 2010. Understanding retweeting behaviors in social networks. In *Proc. of the 19th ACM CIKM*, 1633–1636. ACM.