From Topics to Opinions:
Modelling Subjectivity for Retweeting Analysis on Twitter

Abstract

In this paper, we investigate how user's subjectivity influence their information diffusion behavior. Inspired by 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. We also put forward a new way to measure the subjectivity similarity between two subjectivity models. For the retweeting behavior analysis, three factors (attractiveness, sociality and popularity) are considered based on the subjectivity similarities among a target tweet, its author and followers. In the experiments, we demonstrate that a user is more likely to retweet a message considering the influence of the three factors and the utility of our model in retweeting analysis is verified qualitatively and quantitatively on real Twitter dataset.

Introduction

Information diffusion has drawn considerable research attentions from computer scientists, especially in the area of online social networks. Researchers 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). In this work, we target at a different problem: given a new message, we intend to predict which users will participate in the future diffusion process of this messsage(Bian, Yang, and Chua 2014). An illustration of the problem in a heterogeneous social network of Twitter can be found in Figure 1. In this example, the users have tweeted about two topics: cellphone “Iphone” and movie “Frozen”. Now Tony posts a new tweet about movie “Frozen”, we want to find out which one is more likely to disseminate it among all the receivers of the new tweet.

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

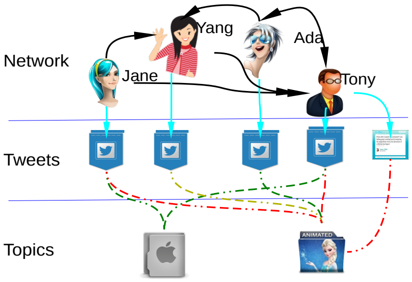

Figure 1: Motivating example. For opinions of different users, the color “red” stands for positive evaluation, “green” for negative, and “yellow” for neutral.

The propagation behaviors are different in different social networks. Twitter plays an important role in the process of information dissemination on the Internet because the retweeting convention provides an unprecedented mechanism for the spread of information despite the restricted length of a single message. Actually almost a quarter of the tweets are retweeted from others (Yang et al. 2010). Understanding how retweeting behavior works can help explaining information diffusion on Twitter.

Previous studies have developed a variety of techniques and models to capture the factors of retweeting behaviors (Macskassy and Michelson 2011; Feng and Wang 2013). However few studies have investigated the subjective motivation of a user to retweet a message. From the point of a user, retweeting is a process that includes following, evaluating and deciding whether to share. The crucial part is to evaluate whether a tweet contains information worthy enough to be shared. Therefore modelling 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 more prone to retweet a message that meets his own tastes. In Figure 1, the tweets of the users present their different opinions about two topics. Tony and Jane were positive about movie “Frozen”, while Ada was negative and Yang was neutral. For the new tweet of Tony which is positive about “Frozen”, Jane is more likely to retweet it because they both like the movie.

For the problem to be investigated, there are two questions arising: how to accurately model the subjectivity of users in terms of topics and opinions, and how to effectively measure the worthiness for the users to retweet? 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 considering his subjectivity.

The rest of the paper is organized as follows: firstly related works are described; we give the definition and establishment details of the proposed subjectivity model, before the subjectivity similarity is defined; then the factors are specified for the retweeting analysis problem;following are experiments of quantitative evaluation; and we summarizes the paper and points out future work finally.

Related Work

A large body of studies have analyzed characteristics of retweeting behavior (Bian, Yang, and Chua 2014; Luo et al. 2013), examining 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 neglect the subjectivity of users, which is the underlying reason for the retweeting behaviors.

Previous researches of sentiment analysis have mainly focused on reviews (Liu 2012). Recently, there have been many works on sentiment analysis for informal social media langage, mainly focusing on the message level (Jiang et al. 2011; Tan et al. 2011; Guerra, Meira, and Cardie 2014). Topic models can also be utilized in sentiment analysis to correlate sentiment with topics. For example, Mei *et al.* (Mei et al. 2007) and Lin *et al.* (Lin and He 2009) attempted to incorporate the sentiment factor into topic models. Usually they learn a general word-sentiment distribution to model the sentiment of blogs or reviews, which may not work well for short and informal social media languages. Sentiment expression is deemed to be more challenging as sentiment is often embodied in subtle linguistic mechanisms such as: negation, capitalization, repeated letters, exclamation and emoticon(e.g. "happy!!"), intensifiers (e.g. "liked" versus "liked verymuch") and diminishers (e.g. "excellent" versus "rather excellent")etc (Brody and Diakopoulos 2011). These are hard to be modeled with probabilistic distribution. However, rule-based sentiment analysis methods can catch such subtle sentiment expressions by transforming the them into rules (Thelwall et al. 2010). In our model establishment framework, we adopt rule-based method for tweet sentiment analysis.

Subjectivity Model

Subjectivity has been extensively studied by psychologists to characterize the personality of a person based on his historical behaviors and remarks (Engbert et al. 2007). Linguists define the subjectivity of language as speakers always show their perspectives, attitudes and sentiments to events, people, topics, and entities in their linguistic contents (Stein and Wright 2005). In the computer sicience, opinion mining techniques (Liu 2012) have been developed to computationally model the subjectivity of users. With the explosion of social media over the past decade, more and more User-Generated Content (UGC) is available on the Web for expressing users' opinions. A variety of aspect-level and topic-sentiment models have been built from UGC by casting opinions as polarity, strength, or emotions regarding a topic (Lek and Poo 2013; Mei et al. 2007). We give a general framework to model subjectivity by combining topics and opinions together. Here we give our definition of subjective model under context of Twitter, while we emphasize that our model can be transfered to other data platforms 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 topics users in V talk about, and a sentiment space S to evaluate their opinions towards 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 $\{O_k\}$ towards each topic distributed over sentiment space S.*

$$SM_u = \{(k, w_{u,k}, \{d_{u,k,s} | s \in S\}) \, | k \in T\} \qquad (1)$$

where:

- *with respect to user u, for each topic $k \in T$, its 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 the user towards topic t is modelled as a topic-dependent sentiment distribution over sentiment space S, $O_k = \{d_{u,k,s} | s \in S\}$, subject to $\sum_{s=1}^{|S|} d_{u,k,s} = 1$.*

Our model is more general than others in that we combine the topics of interest and topic-dependent opinions into a holistic framework, and more importantly, we define opinion as a probabilistic distribution in a scalable sentiment space. The sentiment space can cover all the sentiment modalities, for example, it could be binary value standing for positive and negative polarity, or sequential value for sentiment strength, or discrete value for various emotions. Figure 2 is a visualized subjectivity model example in a $[0, 100]$ topic space and a discrete $[0, 8]$ sentiment space. The definition of the subjectivity model is in an abstract form, which needs to be concreted from two aspect:(1)how to construct the subjectivity model (2)how to utilize this model for retweeting analysis.

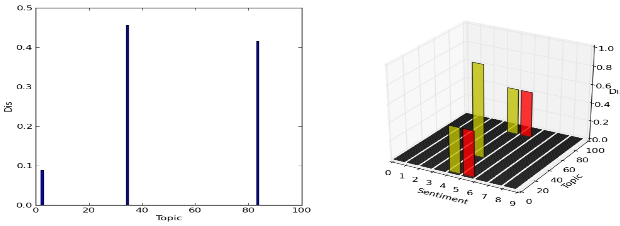

Figure 2: Subjectivity model example. The left subgraph denotes interests distribution on topic 2, 32 and 83: $(w_{u,2} = 0.08, w_{u,32} = 0.48, w_{u,83} = 0.44)$. The right subgraph denotes opinions towards topics: $O_2 = (d_{u,2,4} = 0.5, d_{u,2,5} = 0.5)$, $O_{32} = (d_{u,32,4} = 1.0)$, $O_{83} =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

Establishment of Subjectivity Model

In this section, we present our method to construct subjectivity model SM_u by deriving topics and opinions from the message histories of all users $M = \{M_u | u \in V\}$.

Topic Analysis In this paper, we simply use the concept of topics to broadly refer to the different kinds of content such as bag-of-words, entities, hashtag, etc. 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). State-of-the-art topic model such as Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) is employed for unsupervised topic discovery and for topic assignment of future documents. 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 regards all tweets of a user as one document of LDA (Hong and Davison 2010). The TM model will be used throughout our framework. For a single tweet t, each dimension i of its topic distribution θ_t can be obtained as follows:

$$\theta_{t,i} = \frac{\prod_{w \in t} P(w|\phi_i)}{\prod_{w \in t} \sum_k P(w|\phi_k)} \tag{2}$$

Usually a tweet concentrates a single topic within its short length, therefore we assign a tweet t to a topic that maximizes the probability of generating t:

$$z_t = \arg\max_k \prod_{w \in t} P(w|\phi_k) \tag{3}$$

Finally we can get the weight distribution on each topic k of user u by normalizing all tweets that talk about topic k:

$$w_{u,k} = \frac{|\{t : t \in M_u \wedge z_t = k\}|}{|M_u|} \tag{4}$$

Opinion Analysis Most previous opinion mining researches (Liu 2012) represent opinion with a porlarity label in a binary sentiment space (0, 1): 1 means opinion agreement or positive sentiment, 0 means opinion disagreement or negative sentiment. However, such approach cannot fully distinguish user's detail opinion. For example, in Figure 1, both Tony and Jane holded overall positive opinion about the movie "Frozen". Jane mainly liked the romantic story of this movie but was a little disappointed about its animation picture, while Tony liked this movie because he was mostly convinced by its animation technology although he disliked the prince and princess genre. If we represent their opinions with a simple binary polarity manner, without differentiating their preferences of different aspects, the subjectivity model may not satisfy the information discovery needs of different users. Therefore,considering different aspects of a topic, 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 the 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 is got as:

$$\begin{aligned} O_k &= \{d_{u,k,s} | s \in S\} && (5) \\ &= \{\frac{|t : t \in M_u \wedge z_t = k \wedge s_t = s|}{|M_u|} | s \in S\} && (6) \end{aligned}$$

Subjectivity Similarity

With the subjectivity models estabilshed, the subjectivity similarity needs to be defined and calculated to simulate the various subjective decision-making process such as retweeting behavior. In this section, we define a novel similarity measurement to quantify the subjectivity similarity.

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. At the same time, values in the sentiment space are not independent. They are sequential in magnitude and quantified to measure the strength of sentiment. Therefore, normal distrbution similarity measurements such as KL-divergence and consine similarity are not suitable for such kind of opinion distribution. For example,illustrated in Table 1, in a $[0, 8]$ integer sentiment space, opinion O_k^1 is most negative (100% of value 0), opinion O_k^2 (100% of value 6) is positive, and O_k^3 (100% of value 8) is most positive. If the cosine similarity measurement is adopted to calculate opinion similarity, all similarities among them are 0. In fact O_k^2 is more similar with O_k^3 than O_k^1 because they both hold positive opinion and their sentiment strength distance is much less than with O_k^1. Therefore, opinion similarity can't be calculated simply as the normal probabilistic distributions, or just as the sentiment strength distance. To accurately catch opinion similarity, we propose a novel method by combining both sentiment distance and distribution similarity. The opinion similarity between two opinions O_k^1, O_k^2on the same topic k can be calculated as:

$$Sim(O_k^1, O_k^2) = \frac{|S| - |\sum_{i=0}^{|S|} d_i^1 v_i - \sum_{i=0}^{|S|} d_i^2 v_i|}{|S|} \quad (7)$$

where d_i denotes the i^{th} dimension of opinion distribution, and v_i denotes corresponding sentiment strength value. The similarities of opinions in Table 1 calculated with Equation 7 are $Sim(O_k^1, O_k^3) = 0$, $Sim(O_k^2, O_k^3) = 6/8$ and $Sim(O_k^1, O_k^2) = 1/8$, which are consistent with our intuitive understanding.

Table 1: Illustration of opinion similarity

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

Subjectivity Similarity As the subjectivity model indicates, a user may be interested in several topics and the weights of interests is a distribution over topic space T.Therefore, the subjectivity similarity between two subectivity models should be integrated from the distribution of topics of interest and the opinion similarities on each topic. Accordingly, overall sbjectivity similarity between two subjectivity models SM_u, SM_v can be calculated as:

$$Sim(SM_u, SM_v) = \frac{\sum_{k=1}^{|T_{u,v}|} \theta_u(k) Sim(O_k^u, O_u^v)}{\sum_{k=1}^{|T_{u,v}|} \theta_u(k)} \quad (8)$$

where $T_{u,v}$ denotes the common topics between two users u and v, which can be regarded as the intersection between their topics of interest with $\theta_u^k \neq 0 \wedge \theta_v^k \neq 0$; $\theta_u(k)$ denotes the topic k weight of user u, and $\sum_{k=1}^{|T_{u,v}|} \theta_u(k)$ is the normalized factor.

Note that, the subjectivity similarity is asymmetric because when we measure how similar user u is with user v we use the weights of common interests of user u. The intuition lies in that subjectivity of a user is a personal inner interest and taste, and it is also a subjective decision to judge a like-minded person for the user. For our measurement of subjectivity similarity, $Sim(SM_u, SM_v) \neq Sim(SM_u, SM_v)$.

Retweeting Analysis

Whether a user propagate a tweet may be affected by various factors. From the point of a user, a tweet is more likely to be retweeted because the user finds its content worth to. Generally three factors need to be considered: 1) the content of the tweet is in accordance with a user's subjectivity, which are attractive for the user to retweet; 2) the tweet is posted by the user's like-minded close friend, and his retweeting behavior is due to social needs; and 3) the content is novel or epidemic, and his retweeting behavior is a result of conformity needs (i.e., the act of matching opinions to group norms (Cialdini and Goldstein 2004)). In this section, we firstly formulate the retweeting analysis as a classfication problem, then present the analysis of the three factors considering subjectivity for understanding the underlying reasons that a user disseminates the tweet based on his subjective initiative.

Problem Formulation

For a target tweet m, let F denote the followers who receive m by following its author u_a, and for each user $u \in F \cup \{u_a\}$, let M_u denote a tweet collection u has published. For each follower $f \in F$, we can define a quadruple $< f, u_a, m, r_f >$:

- r_f is a binary label indicating if m is retweeted by u, which should be predicted for a future tweet.
- Firstly our work focuses on building subjectivity model $P(u)$ for each user $u \in F \cup \{u_a\}$ with all tweets collections $M = \{M_u | u \in F \cup \{u_a\}\}$.
- Then we investigate the relation between the subjectivity of a user and his retweeting behavior to predict r_f by calculating subjectivity similarities between tweet m, its author u_a and follower u.

Retweeting Analysis

The three factors exhibit different types of reasons a user rewteet a microblog, which will be analyzed and quantized in this section.

Attractiveness A user is likely to repost a microblog 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 3, and let s_t be its sentiment. The content of t can also be modeled using our subjectivity model definition with a single topic of interest and a %100 opinion distribution on a single sentiment value. Thus the attractiveness of the tweet can be measured with subjectivity similarity using Equation 8, 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 t 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 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 tweets from Tony about movie, but may not be 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 calculates the relative expertise of the tweet author among his followers including f. The expert user impose more influence on others than the others do. 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 Twit-

ter[1].

Similarity Factor $w_S(u_a, f)$: The similarity of interests of u_a and f is measured as the inverse KL-divergence between their topic 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:

$$w_{u_a,f} = \lambda_1 * w_E(u_a) + \lambda_2 * w_L(u_a) + \lambda_3 * w_S(u_a, f) + \lambda_4 * w_I(u_a, f). \tag{9}$$

where λ_i is an optional weight vector to enable different influence of the factors, subject to $\sum_{i=1}^{4} \lambda_i = 1$.

Popularity If a tweet is popular (e.g., breaking news), it will be very probable to be retweeted. In this situation, the tweet t is often inconsistent with the interests and opinion of its author u_a. Thus the similarity between t and u_a in terms of subjectivity is very 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 w_P 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 and it is popular among his friends. In next section we carry out a set of experiments to inspect and verify the impact of such motivation on retweeting behavior.

Experiments

Dataset and Settings

We adopt the Twitter dataset of a previous work (Luo et al. 2013). To form the dataset, 500 target English tweets published from September 14th, 2012 to October 1st, 2012 were monitored to find who would retweet it in the next days. Besides, each target tweet was set as starting point to collect at least 200 historical tweets for its author and followers. Overall, there are 45,531 users who have posted at least 6,277,736 historical tweets, 5214 of which retweet at least one target tweet during the monitored period. To avoid the bias introduced by dataset imbalance, an evaluation dataset was constructed by taking 5,214 retweeters as positive instances, and randomly sampling 5,214 non-retweeters as negative instances.

For the topic model of LDA, we use variational inference-based topic model package Gensim (Řehůřek and Sojka 2010), which adopts an efficient batch-based online inference algorithm and can easily adapt to new document. All parameters are set as defaults and the number of topic traverses from 50 to 200. For sentiment analysis of each tweet, 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 whether 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 model.

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 as Figure 3. The right part of each sub-figure illustrates topic distribution and the left part illustrates opinions towards each topic. Figure 4 shows top words of the 14^{th} topic, the tweets of the author and two followers in word cloud diagrams[2].

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 about two topics (the 14^{th} and 52^{nd} topic) uniformly and his opinion towards 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.

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

[2] We use TagCrowd (http://tagcrowd.com/) to produce word cloud.

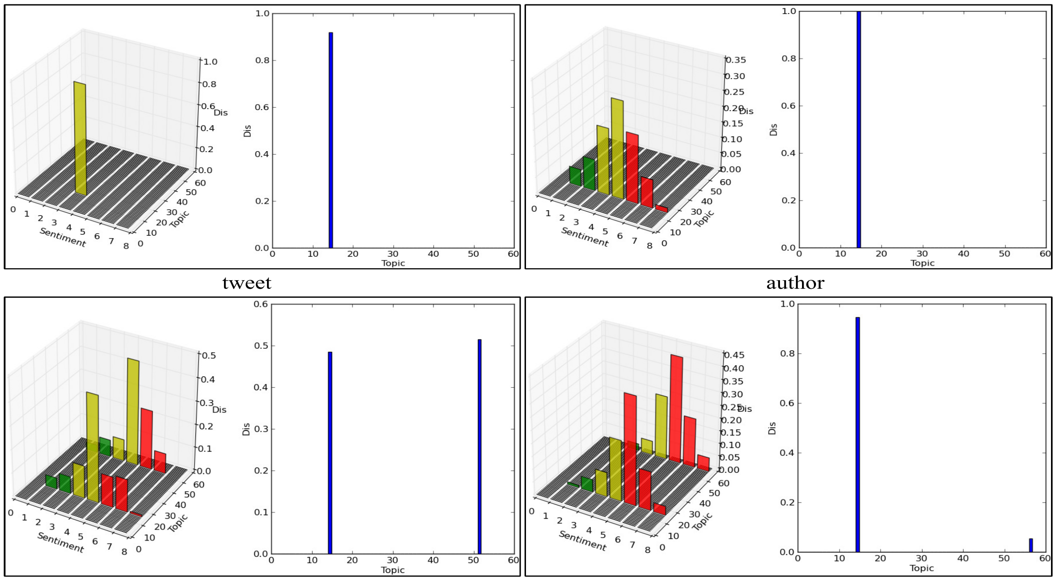

Figure 3: An illustration of subjectivity models of a tweet, author and two followers.

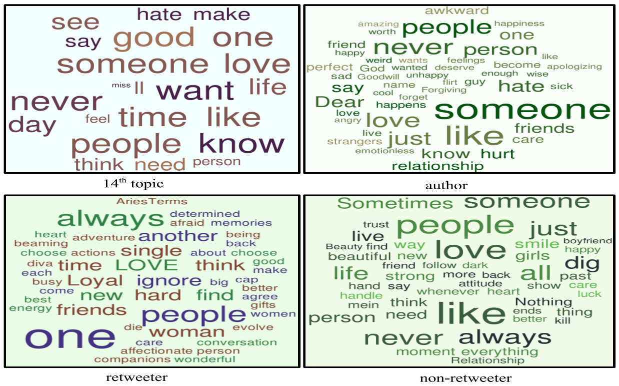
see hate make say good one someone love never miss ll want life day feel time like people know think need person
awkward amazing people happiness worth one friend never person happy like perfect God wanted deserve become apologizing sad Goodwill unhappy enough wise say name flirt guy hate sick Dear cool forget Forgiving happens someone love angry love live friends strangers just like care emotionless know hurt relationship
AriesTerms always determined afraid memories heart adventure another being beaming back choose actions single about choose diva time LOVE think good each make busy Loyal ignore big cap come better best new hard find agree energy friends gifts people women one die woman evolve care conversation affectionate person companions wonderful
Sometimes someone trust live people just boyfriend Beauty find way smile happy beautiful new love girls dig friend follow dark life strong more back all past hand say attitude show care handle whenever heart luck mein think Nothing person need like ends thing better kill never always moment everything Relationship
14th topic
author
retweeter
non-retweeter

Figure 4: Word cloud diagrams of the 14^{th} topic, author and followers.

Table 3 shows the three factors measurement for both retweeter and non-retweeter. Two followers 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 carry out the retweeting behavior prediction experiments with three stages. Firstly we compare our model against other topic-based models including TF-IDF model (modelling user interests with bag-of-words), entity-based model (modelling user interests with entities extracted from the UGC) and hashtag-based model (modelling user interests with hashtags used in the UGC)(Abel et al. 2011).

Secondly, our model is 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 use Equation 8 for the subjectivity similarity of TSM and JST to get three factors values as our model.

Finally, our model tries 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 characteristics of users. Therefore, our model is also compared with the method of Luo *et al.* (2013) (marked as “LUO”), in which diffenent factors that might affect rewteeting behaviors have been considered. In their work they use bag-of-words to model user interests, and we also carry out combining experiments to demonstrate that performance of prediction can be improved by replacing their bag-of-words model with our model (marks with “LUO+” prefix).

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. The classification 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-

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** $*$ ‡

based model, its accuracy is only 59.12%, the reason lies in the spacity of hashtag in tweets.

Secondly, in the comparative results, $Sim(f,t)$ and sim_{all} outperform "LUO" (71.76%) significatantly. The best performance is achieved by the sim_{all} (75.64%), for which we add three factors values to the classifier to test the impact of their combination. The perfromance 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 strength, which could distinguish different subective motivation of retweeting behavior.

Finally, in the combining evaluation experiment, $Sim(f,t)$ gives a significant improvement (LUO+$Sim(f,t)$, 2.12% improvement) over "LUO", but other two factors 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 factors.

The results above show that subjectivity model can better help predicting retweeting behavior than other models and can be regarded as a better way to model the users for retweeting behavior analysis.

Conclusion

Motivated by the psychological research, this paper postulates that the online behaviors of social media users are affected by their subjectivity. Therefore, a general subjectivity model has been proposed by combining topics and opinions to model the subjectivity of the users. Also an framework has been designed to establish the subjectivity model, and a novel method is put forward to measure the subjectivity similarity. The subjectivity model has been applied to the retweeting analysis considering three different factors. 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 the subjectivity 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 *Proceedings of the 37th International ACM SIGIR Conference on Research &#38; Development in Information Retrieval*, 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.

Brody, S., and Diakopoulos, N. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 562–570. Association for Computational Linguistics.

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 *Proceedings of the 23rd international conference on World wide web*, 925–936. International World Wide Web Conferences Steering Committee.

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.

Engbert, K.; Wohlschläger, A.; Thomas, R.; and Haggard, P. 2007. Agency, subjective time, and other minds. *Journal of Experimental Psychology: Human Perception and Performance* 33(6):1261.

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 *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, 443–452. New York, NY, USA: ACM.

Guille, A.; Hacid, H.; Favre, C.; and Zighed, D. A. 2013. Information diffusion in online social networks: A survey. *ACM SIGMOD Record* 42(1):17–28.

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. International World Wide Web Conferences Steering Committee.

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. International World Wide Web Conferences Steering Committee.

Jiang, L.; Yu, M.; Zhou, M.; Liu, X.; and Zhao, T. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 151–160. Association for Computational Linguistics.

Kempe, D.; Kleinberg, J.; and Tardos, É. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 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 on*, 366–373. IEEE.

Lin, C., and He, Y. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, 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 *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 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 *Proceedings of the 16th international conference on World Wide Web*, 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 *Proceedings 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 *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

Stein, D., and Wright, S. 2005. *Subjectivity and Subjectivisation: Linguistic Perspectives*. Cambridge University Press.

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 *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 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.