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

Abstract

Social media such as Twitter provides researchers with abundant User-Generated Content (UGC) for analyzing users' online behaviors. In this paper, we focus on retweeting behavior, which is one of the key mechanisms of information dissemination on Twitter. To understand the motivation of retweeting behavior, previous studies have committed to modelling interests of users with topics derived from UGC, but few have investigated opinions of users. Inspired by psychological research, we propose a novel subjectivity model by combining both topics and opinions articulated in UGC. We also put forward a new way to measure the subjectivity similarity between two subjectivity models, and demonstrate that a user is more likely to retweet a message with approximate subjectivity similarity. In the experiments, the subjectivity similarity is verified to be correlated with retweeting behavior by a statistical hypothesis test. Comparing with other topic-based models in retweeting prediction, our model obtains the best evaluation performance in terms of accuracy. Furthermore the proposed model gives significant accuracy improvement over an off-the-shelf predicting model considering other factors.

Introduction

Microblogging has become a center of attention in the area of social media due to the amount of users it has attracted and the volume of messages it produces. Microblogging services such as Twitter appear to play an important role in the process of information dissemination on the Internet, making it possible for messages to spread virally in a matter of minutes. The retweeting convention and complex network of Twitter provide an unprecedented mechanism for the spread of information despite the restricted length of a single message (i.e. tweet). Actually almost a quarter of the tweets are retweeted from other users (Yang et al. 2010). Understanding how retweeting behavior works can help explaining information dissemination on Twitter.

There have been many studies trying to identify factors that influence whether a tweet will be retweeted (Boyd, Golder, and Lotan 2010; Kwak et al. 2010). However few studies have investigated the subjective motivation of a user to retweet a message. The subjective initiative nature of human determines that his behavior pattern is subjectivity driven. Psychological researchers have identified subjectivity as the underlying factor that influences human's behaviors (Moore and Haggard 2008). Also according to theory of Biased Assimilation, people tend to choose and disseminate information according to their own biased subjectivity (Hyman 2000). Users receive thousands of tweets on different topics every day, whether a tweet will be retweeted will depend on the subjective choice of users. From the point of a user, retweeting is a process that includes reading the tweet, evaluating the content and deciding whether to share. The crucial part is to evaluate whether a tweet contains information interesting to the user who might find that it is worthy to be shared. Therefore modelling the subjective motivation of users will provide an important perspective for retweeting behavior analysis. This research is motivated by a desire to find what drives users of social media to disseminate information they come across.

Previous studies on retweeting analysis have shown that an enriched user model gives coherent and consistent explanation for retweeting analysis (Macskassy and Michelson 2011; Feng and Wang 2013). Specifically, researchers have tried to model users from four types of information: profile features ("**Who you are**"), tweeting behavior ("**How you tweet**"), linguistic content ("**What you tweet**") and social network ("**Whom you connect**") (Pennacchiotti and Popescu 2011). Especially, interests of a user, i.e. topics encapsulated in User-Generated Content (UGC), have been proved consistently dependable for behavior analysis (Petrovic, Osborne, and Lavrenko 2011). However, to our best knowledge, few studies have considered the subjective aspect ("**what's your opinions**") when modelling a user. In this paper, we propose a novel method to model subjectivity of users (defined as subjectivity model) by combining both the topics and opinions.

Users of social media usually present their opinions by gennerating subjective content on topics they are interested in. The subjectivity of a user is encoded in the UGC on Twitter. Therefore, we explore the tweets a user has published to establish the subjectivity model. To meet the challenges of data sparsity and computational complexity, we design an algorithm to build the subjectivity model by making use of the local network structure and homophily of social network. For the retweeting analysis problem, we assume that the probability a user retweets a message could be evaluated

by a subjectivity similarity measurement. Therefore, we put forward a new way to measure the subjectivity similarity, and use three subjectivity similarities among tweets, authors and followers to predict retweeting behavior. Expertiment results show that retweeting behaviors are correlated with all three subjectivity similarities, the subjectivity model outperforms topic-based model for retweeting prediction, and the performance of an off-the-shelf predicting model is significantly improve by combining with our model.

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

Subjectivity Model

Subjectivity has been extensively studied by psychologists to characterize the personality of a person based on his historicalal 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). However, how to computationally model the subjectivity of a user is still an open challenge. The advent of online social media such as Twitter has given a new layout to the challenge. Twitter allows users to show their personal subjectivity by publishing short messages, which provides researchers with data resources to model the subjectivity of users. Therefore, we give a formal definition of the subjectivity model under the context of Twitter.

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 valence 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 $P(u)$ of user u, is the combination of topics $\{t\}$ the user talks about in topic space T and his opinions $\{O_t\}$ towards each topic distributed over sentiment valence space S.*

$$P(u) = \{(t, w_u(t), \{d_{u,t}(s) | s \in S\}) | t \in T\} \quad (1)$$

where:

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

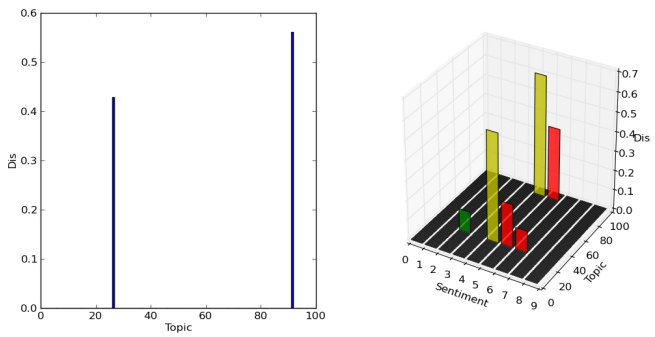

Figure 1: Subjectivity model example. The left subgraph denotes interests distribution on topic 23 and 86: $(w_u(23) = 0.43, w_u(86) = 0.57)$. The right subgraph denotes opinions towards two topics: $O_{23} = (d_{u,23}(2) = 0.1, d_{u,23}(4) = 0.5, d_{u,23}(5) = 0.3, d_{u,23}(6) = 0.1)$, $O_{86} = (d_{u,86}(4) = 0.66, d_{u,86}(5) = 0.34)$.

Figure 1 is a visualized subjectivity model of a user in a $[0, 100]$ topic space and a $[0, 8]$ sentiment valence space.

Specially, the topics and opinions of a tweet can also be represented by the subjectivity model in the form of Equation 1. Therefore we assume that there is also a subjectivity model for a tweet, and will not differentiate it from the subjectivity model of a user in this paper.

Establishment of Subjectivity Model

The definition of the subjectivity model is in an abstract form by using latent concepts of topics and opinions, which need to be derived from the message histories of all users $M = \{M_u | u \in V\}$.

Topic Analysis Topic analysis for all users in a global network on Twitter is a non-trivial task. There are hundreds of millions of users and billions of tweets associated with these users. The effectiveness and efficiency of the topic analysis algorithm is a challenge. However, the follow relationship on Twitter is a strong indicator of a phenomenon called "homophily", which has been observed in many social networks (McPherson, Smith-Lovin, and Cook 2001). Homophily implies that a user follows another user because of sharing common interests. According to the principle of homophily, we put forwards the concept of **local topic space** by combining topic analysis with network topology on Twitter:

Definition 2 (Local Topic Space) *In a global social network $G = (V, E)$, for a user $u \in V$, we use $G_u^\tau \subseteq G$ to denote u's τ-ego network, where τ-ego network means subnetwork formed by u's τ-hop friends in the network G, and $\tau \geqslant 1$ is a tunable integer parameter to control the scale of the ego network. For the τ-ego network of u, all users' interests are assumed concentrate on limited topics derived from their UGC, and these topics form a local topic space T_u.*

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 such as Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) is more effective in identifying topics from short and informal social media language (Hong and Davison 2010). Therefore we adopt the user-level LDA model for topic analysis, which regards all tweets of a user as one document of LDA. The LDA model is adapted to our local topic space assumption, and the relatively tiny size and topic concentration of users in an ego network lower the impact of data sparsity, and degrade the computational difficulty of LDA.

Opinion Mining In the Natural Language Processing domain, opinion mining or sentiment analysis is formally defined as the computational study of sentiments and opinions about topics expressed in a text (Liu 2012). Opinions are often regulated as sequential discrete values to represent sentiment strength. Researches on the sentiment analysis of social media have provided effective techniques and tools (Thelwall et al. 2010; Hu et al. 2013). In this work,we just make use of the off-the-shelf work, i.e. SentiStrength (Thelwall et al. 2010). SentiStrength assigns two values to each tweet standing for sentiment strengths: a negative value within $[-5,-1]$ denoting negative strength, and a positive value within $[1,5]$ denoting positive strength. The $[-5,5]$ sentiment valence space can be used to catch fine opinion distributions in the subjectivity model. For the convenience of calculation, we map the output of SentiStrength to a single value in sentiment valence space $[0,8]$ as follows:

$$o = \begin{cases} p+3 & if|p| > |n| \\ n+5 & \text{if } |n| > |p| \\ 4 & \text{if } |p| = |n| \end{cases} \quad (2)$$

where p denotes the positive strength and n denotes the negative strength.

Concreting Subjectivity Model As Definition 2 describes, a τ-ego network $G_u^\tau = (U, E_u)$ for a user u can be extracted from global network. Then the subjectivity model of each user $u \in U$ can be concreted within the ego network. Let M_u denote tweets collection published by user u, and $M = \{M_u | u \in U\}$ denote all tweets collections of users in G_u^τ. A topic model $P(\theta, \beta | M)$ can be constructed with user-level LDA model, of which the parameter θ reprents user-topic distribution and β reprents topic-vocabulary distribution. All topics of the topic model form a local topic space T_u. The parameter θ_u represents the topic distribution of user u over T_u. Simultaneously SentiStrength is applied to each tweet $m \in M_u$ and outputs sentiment strength s_m. The subjectivity model $P(u)$ is established as follows:

- Step 1, the parameter θ_u naturally corresponds to interests distribution of user u in the local topic space T_u, and the topics u talks about are $Z_u = \{t | p(t|\theta_u(t)) > 0, t \in T_u\}$.
- Step 2, the topic model is applied to each tweet m to identify topics it talks about, denoted as $Z_m = \{t | p(t|\theta, \beta) > 0, t \in T_u\}$.
- Step 3, the opinion distribution of user u towards topic $t \in Z_u$ can be calculated as:

$$O_t = \left\{ d_{u,t}(o) = \frac{N_o}{\sum_{o \in O} N_o} | o \in O, O = [0,8] \right\} \quad (3)$$

where N_o is the number of times user u expresses an opinion towards topic t with sentiment strength o, which can be calculated as:

$$N_o = \sum_{m \in Mu} I(s_m), \text{ if } s_m = o \& t \in Z_m \quad (4)$$

$$I(s_m) = \begin{cases} 1 & \text{if } s_m = o \& t \in Z_m \\ 0 & \text{else} \end{cases} \quad (5)$$

For simplicity, it is postulated that the sentiment of each tweet s_m is related to all topics it talks about in Z_m. As a future work, we will adopt more sophisticated method to identify opinion towards each topic in a tweet.

As a special case, we can also establish a subjectivity model $P(m)$ for a tweet m with only step 2 and 3. Note that the opinion distribution for each topic t of the tweet is $(d_{m,t}(s_m) = 1.0)$.

Retweeting Analysis With Subjectivity Model

Apart from the context constraints such as network topology, a tweet is more likely to be retweeted by a user who finds its content worth to. Therefore, we are not interested in modelling the tweet by itself as other researchers (Naveed et al. 2011; Pfitzner, Garas, and Schweitzer 2012), but understanding the underlying reasons that a user disseminates the tweet based on his subjective initiative. We assume that if a tweet is published by the author, all followers will read it in time. Under such assumption, we investigate the problem within a 1-ego network for the author of target tweet. In the ego network, the relations among a tweet, the author and followers are illustrated as Figure 2.

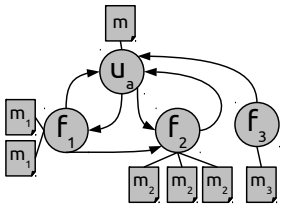

Figure 2: Illustration of relations among tweet, author and followers. Author is denoted as u_a, tweet as m, followers as f_i and tweets of follower f_i as m_i. An directed edge (f_i, u_a) means that f_i is exposed to the messages published by u_a.

Problem Formulation

The retweeting analysis problem can be formulated as following: 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 $u \in F$, we can define a quadruple $< u, u_a, m, r_f >$:

- r_f is a binary label indicating if m is retweeted by u.
- Firstly our work focuses on building subjectivity model $P(u)$ for each user $u \in F \cup \{u_a\}$ in the ego network 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.

Subjectivity Similarity

It is assumed that if a tweet is similar enough with the subjectivity of a user in terms of topics and opinions, the user will have a very high probability to decide to retweet it. With the subjectivity models estabilshed for the users and tweet, the subjective decision-making process can be simulated by calculating the subjectivity similarity between the tweet and users. In this section, we define a novel similarity measurement to quantify the subjectivity similarity, which consists of topic similarity and opinion similarity.

Topic Similarity The similarity between two topic distributions can be calculated with methods such as the cosine distance (Cha 2007) or the Jensen-Shannon Divergence (Weng et al. 2010). We adopt the cosine distance to measure the topic similarity because it performs better than other measurements in our research settings. It is defined as:

$$sim_{topic} = \frac{\theta_m \cdot \theta_u}{\parallel \theta_m \parallel \parallel \theta_u \parallel} \tag{6}$$

where θ_u denotes the topic distribution of user u and θ_m denotes the topic distribution of tweet m.

Opinion Similarity Opinion in the subjectivity model is treated as a distribution over sentiment valence space with each entry of the distribution representing the proportion of the corresponding value in the overall sentiment values. However, values in the sentiment valence space are not independent. They are sequential and represent strength of the sentiment. Illustrated as Table 1, opinion O_t^1 is the most negative towards topic t (100% of strength value 0), while opinion O_t^2 (100% of strength value 7) and O_t^3 (100% of strength value 8) are both positive. If the cosine similarity measurement is adopted to calculate opinion similarity, all similarities among them are 0. In fact O_t^2 is more similar with O_t^3 than O_t^1 because they both hold positive opinion and their sentiment distance is much less than O_t^1. Therefore, opinion similarity can't be calculated simply as the topic distributions. To accurately catch opinion similarity, we propose a novel method by combining both sentiment distance and distribution similarity. The opinion similarity between two opinions on the same topic t can be calculated as:

$$sim^t_{opinion}(O_t^1, O_t^2) = \frac{8 - |\sum_{i=0}^{8} d_i^1 v_i - \sum_{i=0}^{8} d_i^2 v_i|}{8} \tag{7}$$

where d_i denotes the i^{th} entry of opinion distribution vector, and v_i denotes corresponding sentiment strength value. The similarities of opinions in Table 1 calculated with Equation 7 are $sim(O_t^1, O_t^3) = 0$, $sim(O_t^2, O_t^3) = 7/8$ and $sim(O_t^1, O_t^2) = 1/8$, which are consistent with 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	0.0	1.0	0.0
O_t^3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

Accordingly, overall opinion similarity between two subjectivity models can be calculated as normalized similarity of all opinion similarities on common topics.

$$sim_{opinion}(u_1, u_2) = \frac{\sum_{t=1}^{|T|} sim^t_{opinion}(O_t^1, O_t^2)}{|T|} \tag{8}$$

where T denotes the common topics between two subjectivity models, which can be regarded as the intersection between their topic sets Z_{u_1} and Z_{u_2} described in the section of subjectivity model establishment.

Subjectivity Similarity By combining topic similarity and opinion similarity, the subjectivity similarity can be defined as follows:

$$Sim_{sub}(m, u) = \lambda * sim_{topic} + (1 - \lambda) * sim_{opinion} \tag{9}$$

where λ is the coefficient used to control the proportions of topic similarity and opinion similarity in the holistic subjectivity similarity. A user cares more about topics with a larger λ, and cares more about opinions with a samller λ. A personalized λ can be learned from the retweeting history of a user, which enable us to catch subtle retweeting habit and improve retweeting prediction performance for each user.

Retweeting Analysis

The motivation of retweeting behavior is complicated, which involves the target tweet, its author and followers who is following its author, with their relations illustrated as Figure 2. The idea behind this work is that taking opinions towards interests into account can yield benefits in explaining the subjective motivation of retweeting behavior. Specifically, given a tweet m, the author u_a and any one of the followers u, we consider the probability of user u to rewteet m from three aspects: (i) how similar is the tweet m to the subjectivity of user u in terms of topics and opinions, i.e. $sim_{sub}(m, u)$, (ii) how like-minded are the author u_a and user u considering their similarity of subjectivity, i.e. $sim_{sub}(u_a, u)$, and (iii) how original is the tweet m judged from its similarity with the subjectivity of its author u_a, i.e. $sim_{sub}(m, u_a)$. 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 original from inner subjectivity of its author. 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 previous work (Luo et al. 2013). To form the dataset, 500 target English tweets pub-

lished 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 3,0876 users who have retweeted at least 20 times in their 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. All users in the evaluation dataset were separated into the 1-ego network of their target tweet's author to establish their subjectivity model. For subjectivity similarity, a *mini-batch gradient descent* algorithm was implemented to optimize the coefficient λ in Equation 9 for each user with his retweeting history. Therefore, all λs of three subjectivity similarities ($sim_{sub}(m,u)$, $sim_{sub}(u_a,u)$, $sim_{sub}(m,u_a)$) were optimized to reflect the personalized retweeting habit. As a result, the optimized λs are used to calculate three subjectivity similarities for each user of the evaluation dataset with their own target tweets, which are used to study their retweeting behaviors.

Correlation Test

First of all we want to assess the existence of a correlation between subjectivity similarity and retweeting behavior. To verify such correlation, a statistical hypothesis test called Analysis of Variance (ANOVA) (Fisher et al. 1970) is used. ANOVA tests the *null hypothesis* that samples in two or more groups are derived from the same population by estimating the variance of their means. This test fits our goal of testing whether the retweeters and non-retweeters have the same subjectivity similarity means. ANOVA test produces two output values: the *F-ratio* and the *p-value*. If the difference between the means is due to chance, the expected value of the *F-ratio* is 1.00, otherwise it is larger than 1.00. If the p-value is lower than the significance level α, the *null hypothesis* is rejected, which means the results is considered statistically significant. The significance level is conventionally used at 0.01. At the same time, we carry out the test by varying the topic number of LDA for topic analysis as 50, 100, 150 and 200 to determine the impact of topic number. The results are listed in Table 2, The bold-faced entries mean that the *p-value* is lower than significance level $\alpha = 0.01$.

Table 2: ANOVA results for subjectivity similarities

Similarity		$sim_{sub}(m,u)$	$sim_{sub}(u_a,u)$	$sim_{sub}(m,u_a)$
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

Note that for the topic numbers of 100 and 150, all similarities yield *p-values* below α 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.

Performance Evaluation

To evaluate the performance of retweeting behavior prediction, we firstly 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). The cosine distance is used as similarity measurement for these models as topic similarity in our model for comparison.

In addition, subjectivity model tries to catch the subjective motivation of users based on their UGC, whereas other important factors associated with retweeting behavior are not considered, such as network topology and metadata 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 four feature families: "Retweet History"(follower who have retweeted a user before is likely to retweet again), "Follower Status"(the number of tweets, followers, friends, listed times and verified state), "Follower Active Time"(interaction with other users) and "Follower Interests"(TF-IDF bag-of-words model for user interests). Based on the results of Comparative experiment, we also carry out combining experiments to demonstrate that performance of their method can be improved by using our model instead of bag-of-words model.

Table 3: Accuracy performance. A significant improvement over baseline with $*$ and LUO' model with ‡ ($p < 0.05$).

Feature	Accuracy(%)	Feature	Accuracy(%)
RB	60.85		
TF-IDF	62.85 $*$	LUO	71.76 $*$
entity	68.76 $*$	LUO+entity	72.15 $*$
hashtag	59.12	LUO+hashtag	68.44 $*$
$sim_{sub}(m,u)$	73.88 $*$ ‡	LUO+$sim_{sub}(m,u)$	74.04 $*$ ‡
$sim_{sub}(u_a,u)$	70.04 $*$	LUO+$sim_{sub}(u_a,u)$	70.27 $*$
$sim_{sub}(m,u_a)$	69.64 $*$	LUO+$sim_{sub}(m,u_a)$	71.86 $*$
sim_{all}	**75.64** $*$ ‡	LUO+sim_{all}	**78.15** $*$ ‡

The evaluation dataset is randomly divided into five parts for 5-fold cross-validation. The logistic regression classifier of Scikit-learn machine learning package (Pedregosa et al. 2011) is used for training and testing. It is noted that followers who previously had a history of retweeting might do this in the future, so we set a baseline (marked as "RB"), which simply predicts users who have retweeted the author previously as the retweeters of target tweet. The accuracy is taken as our evaluation metric, and the results are listed in Table 3, in which the compariative results are listed in the left part and the combining results in the right part.

Firstly, all models except the hashtag-based model outperform the baseline (60.85%) significantly. While for hashtag-based model, its accuracy is the lowest (59.12%), the reason might lie in a very low usage of hashtag in a user's tweets.

Secondly, in the comparative results, $sim_{sub}(m,u)$ and sim_{all} outperform "LUO" (71.76%) significatantly. The best performance is achieved by the sim_{all} (75.64%), for which we feed all three subjectivity similarities into the logistic classifier to test the impact of their combination. The perfromance of TF-IDF model (62.85%) is only better than baseline. The entity-based model (68.76%) is very close to $sim_{sub}(u_a,u)$ (70.04%) and $sim_{sub}(m,u_a)$ (69.64%), and the difference is not significant.

Finally, in the combining evaluation experiment, for which the TF-TDF model of "LUO" feature set is replaced with other models, the results are diverse. $sim_{sub}(m,u)$ gives a significant improvement (LUO+$sim_{sub}(m,u)$, 2.12% improvement) over "LUO", but other two subjectivity similarities and the entity-based model can not improve performance significantly. The performance is even degraded after combining with the hashtag-based model. But noticing that, the most significant improvement(LUO+sim_{all}, 6.39% improvement) is achieved by combining with all subjectivity similarities.

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.

Related Work

User modelling provides insights into user's online behaviors. Hannon *et al.* (2010) proposed that users can be modeled by tweets contents and the relation of social networks, while content-based model can find similar users who are "distant" without following relations. Macskassy and Michelson (2011) discovered user interests by leveraging Wikipedia as external knowledge to determine a common set of high-level categories that covers entities in tweets. Ramage *et al.* (2010) made use of topic models to analyze tweet contents at the level of individual, showing improved performance on tasks such as post filtering and user recommendation. Xu *et al.* (2012) proposed a mixture model which incorporated three important factors, namely breaking news, friends' timeline and user interests, to explain user posting behavior. Pennacchiotti and Popescu (2011) proposed a most comprehensive method to model Twitter users for user classification, confirming the value of in-depth features by exploiting the UGC. A large body of studies have analyzed characteristics of retweeting (Macskassy and Michelson 2011; 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 (Petrovic, Osborne, and Lavrenko 2011; 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 decision-making of subjective initiative individuals. We have fisrtly proposed a novel subjectivity model for retweeting behavior analysis.

Conclusion

In this paper, from the point of motivation, we postulate that the online behaviors of social media users are affected by their subjectivity. Therefore, a novel subjectivity model has been proposed by combining topics and opinions to model the subjectivity of users. Also an algorithm has been designed to establish the subjectivity model. To make the algorithm more efficiently, only the users of an ego network are considered and a local topic space is proposed according to the homophily principle. A novel subjectivity similarity measurement has been put forward in terms of topic similarity and opinion similarity. The subjectivity model is applied to the retweeting analysis with three subjecivity similarities among tweets, authors and followers. Experiment results demonstrate the effectiveness of the proposed model in the retweeting analysis problem and show that subjectivity 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 friend recommendation.

References

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

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

Boyd, D.; Golder, S.; and Lotan, G. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, 1–10. IEEE.

Cha, S.-H. 2007. Comprehensive survey on distance/similarity measures between probability density functions. *City* 1(2):1.

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.

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 sixth ACM international conference on Web search and data mining*, 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.

Hannon, J.; Bennett, M.; and Smyth, B. 2010. Recommending twitter users to follow using content and collaborative filtering approaches. In *Proc. of the fourth ACM conference on Recommender systems*, 199–206. 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 international conference on World Wide Web*, 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 international conference on World Wide Web companion*, 657–664. International World Wide Web Conferences Steering Committee.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *Proc. of the 19th international conference on World wide web*, 591–600. 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 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*.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 415–444.

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

Naveed, N.; Gottron, T.; Kunegis, J.; and Alhadi, A. C. 2011. Searching microblogs: coping with sparsity and document quality. In *Proc. of the 20th ACM international conference on Information and knowledge management*, 183–188. ACM.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Pennacchiotti, M., and Popescu, A.-M. 2011. A machine learning approach to twitter user classification. In *ICWSM*.

Petrovic, S.; Osborne, M.; and Lavrenko, V. 2011. Rt to win! predicting message propagation in twitter. In *ICWSM*.

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

Ramage, D.; Dumais, S.; and Liebling, D. 2010. Characterizing microblogs with topic models. In *ICWSM*.

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 *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, 177–184. IEEE.

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.

Weng, J.; Lim, E.-P.; Jiang, J.; and He, Q. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *Proc. of the third ACM international conference on Web search and data mining*, 261–270. ACM.

Xu, Z.; Zhang, Y.; Wu, Y.; and Yang, Q. 2012. Modeling user posting behavior on social media. In *Proc. of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 545–554. ACM.

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 international conference on Information and knowledge management*, 1633–1636. ACM.