From Interests 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 considered 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.

Categories and Subject Descriptors
H.4 [**Information Systems Applications**]: Miscellaneous; H.3.3 [**Information Search and Retrieval**]: Information filtering—*performance measures*

General Terms
Model, Experimentation

Keywords
Twitter, subjectivity, retweet, LDA, sentiment analysis

1. 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 [39]. 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 [4, 19]. 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 [24]. Also according to theory of Biased Assimilation, people tend to choose and disseminate information according to their own biased subjectivity [16]. 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 [22, 10]. 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**") [28]. Especially, interests of a user, i.e. topics encapsulated in User-Generated Content (UGC), have been proved consistently dependable for behavior analysis [29]. 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 and tweets as well (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. Experiment 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.

Our work aims to define and establish the subjective model and identify the role of subjectivity in the processes of information diffusion on Twitter. Our contributions can be summarized as follows:

- In the light of psychological theory, we firstly put forward formal definition of subjective model for users and tweets which model both the topics and opinions simultaneously.
- Based on the state-of-the-art topic model and sentiment analysis techniques, we build subjective model from user-generated content on Twitter and apply it to the retweeting behavior analysis problem.
- We systematically evaluate the impact of subjective model on retweeting behavior and demonstrate that it outperforms other models in rewteeting prediction and gives the most significant improvement over a off-the-shelf predicting 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.

2. RELATED WORK

In this section, we give an introduction to three lines of relevant research work: 1) retweeting analysis, 2)user modelling, and 3) sentiment analysis.

2.1 Retweeting Analysis

A large body of studies have analyzed characteristics of retweeting, examining factors that lead to increased retweetability and designing models to estimate the probability of being retweeted.

As for factors influencing retweetability, Suh *et al.* [34] found that tweets with URLs and hashtags were more likely to be retweeted, and there was a strong linear relationship between the number of followers and the likelihood that the tweet be retweeted. Macskassy and Michelson [22] studied a set of Twitter users over a period of a month and found that models derived from tweet content could explain most of retweeting behaviors. Comarela *et al.* [8] found previous response to the tweeter, the tweeters' sending rate, the freshness of information, the length of tweet could affect followers' response to retweet. Starbird and Palen [32] addressed specifically the retweeting mechanism during crises and found that tweets with topical keywords were more likely to be retweeted.

There were also many works extending the analysis to build retweeting prediction model. Osborne and Lavrenko [29] introduced features such as novelty of a tweet and the number of times the author is listed to train a model with a passive aggressive algorithm, and found the dominance of social features, while tweet features added a substantial boost to the performance. Jenders *et al.* [17] analyzed the "obvious" and "latent" features from structural, content-based, and sentimental aspects of both tweets and users, with respect to their impact on the spread of tweets. They found a combination of features covering all aspects was the key to high prediction quality. Naveed *et al.* [26, 25] introduced interestingness as static quality measure to capture the static content quality of tweets, and quantified it based on such features as emoticons, sentiments and topics a tweet contains, then trained a logistic regression model to predict the probability of retweet for an individual tweet. Feng and Wang [10] built a graph made up of users, publishers and tweets nodes with all sources of information incorporating into nodes and edges, and proposed a feature-aware factorization model to rerank the tweets according to their probability of being retweeted. Pfitzner *et al.* [30] proposed a new measure called emotional divergence to evaluate the retweet probability of a tweet and showed that highly emotional diverse tweets can have up to almost five times higher chances of being retweeted.

From a global perspective, all papers introduced above tried to answer the question of "Whether and why a tweet will be retweeted by anyone?". But they are weak to capture "Whether a tweet is retweetable from a user-centric perspective considering the interests and opinions of users". In this paper, we will try to answer this question by building a subjective model which can capture both the interests and opinions of users.

2.2 User Modelling

With the popularity of social media, researchers have begun to pay close attention to the massive amount of data generated by users, and put forwards several techniques to model users on the data. These studies provide researchers with insights into user online behaviors.

Hannon *et al.* [12] proposed that Twitter users can be modeled by tweets content and the relation of Twitter social network. They found that content-based approach could find similar users who are "distant" without follow relations based on interests extracted from the content of tweets. Macskassy and Michelson [22] discover user's topics of interest by leveraging Wikipedia as external knowledge to determine a common set of high-level categories that covers entities in tweets. Ramage *et al.* [31] made use of topic models to analyze Twitter content at the level of individual users with 4S dimensions, showing improved performance on tasks such as post filtering and user recommendation. These efforts of user modelling on Twitter have simply built model for each user by extracting keywords, entities, categories or latent topics from tweet content.

Some researchers argued that user behavior could easily be affected by some external factors other than user interest. Xu *et al.* [38] proposed a mixture model which incorporated three important factors, namely breaking news, friends' timeline and user interest, to explain user posting behavior. Pennacchiotti and Popescu [28] proposed a most comprehensive method to model Twitter user for user classification. They focused on richer feature sets and confirmed the value of in-depth features by exploiting the user-generated content, which reflect a deeper understanding of the Twitter user and the user network structure.

As introduced in Section 1, previous researches have tried to model users from four types of information: profile features, tweeting behavior, linguistic content and social network. Some studies perceived that the implicit features articulated in the user-generated content play an important role in user behavior analysis, and they have proposed various techniques to capture such in-depth features to model user's interest. Additionally, a few of work identified the correlation between sentiment of users and their behaviors, but they all ignored modelling subjectivity of a user. Motivated by the observation, we firstly put forward subjective model to combine both interests and opinions to model a user.

2.3 Sentiment Analysis

Sentiment analysis is a popular research area for many years. Previous research mainly focused on reviews or news comments. Recently, researchers began to pay more and more attention to social media such as Twitter.

Hu *et al.* [14] interpreted emotional signals available in social media data for unsupervised sentiment analysis by providing a unified way to model two main categories of emotional signals: emotion indication and emotion correlation. Jiang *et al.* [18] focused on target-dependent Twitter sentiment classification, they proposed a method to improve target-dependent Twitter sentiment classification by taking target-dependent features and related tweets into consideration. Asiaee T. *et al.* [2] presented a cascaded classifier framework for per-tweet sentiment analysis by extracting tweets about a desired target subject, separating tweets with sentiment, and setting apart positive from negative tweets. Hu emphet al. [15] extracted sentiment relations between tweets based on social theories, and proposed a novel sociological approach to utilize sentiment relations between messages to facilitate sentiment classification and effectively handle noisy Twitter data. Motivated by sociological theories that humans tend to have consistently biased opinions, Calais Guerra *et al.* [5] addressed challenges of topic-based real-time sentiment analysis by proposing a novel transfer learning approach with a suitable source task of opinion holder bias prediction. Thelwall *et al.* [36, 35] designed SentiStrength, an algorithm for extracting sentiment strength from informal English text by exploiting the grammar and spelling styles in typical social media text. In this paper, we adopt SentiStrength for sentiment analysis to build our subjective model, as a finer grain sentiment strength could give us more detailed opinion of users than binary polarized sentiment.

3. SUBJECTIVITY MODEL

Subjectivity has been extensively studied by psychologists to characterize the personality of a person based on his historicaal behaviors and remarks [9]. 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 [33]. 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.

3.1 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.

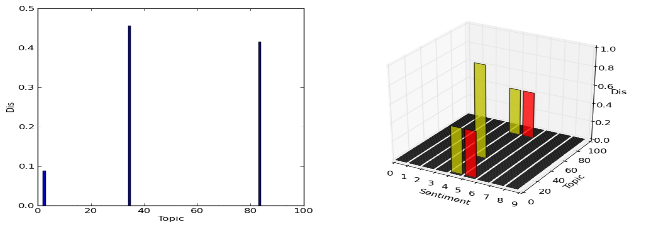

Figure 1: Subjectivity model example. The left subgraph denotes interests distribution on topic 2, 32 and 83: $(w_u(2) = 0.08, w_u(32) = 0.48, w_u(83) = 0.44)$**. The right subgraph denotes opinions towards topics:** $O_2 = (d_{u,2}(4) = 0.5, d_{u,2}(5) = 0.5)$**,** $O_{32} = (d_{u,32}(4) = 1.0)$**,** $O_{83} = (d_{u,83}(4) = 0.5, d_{u,83}(5) = 0.5)$**.**

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

Specially, the content of a tweet can also be represented with the subjectivity model because the topics and opinions of a tweet can be modeled as Equation 1.

3.2 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 $M = \{M_u | u \in V\}$.

3.2.1 Topic Analysis

Topic analysis for all users in a global network on Twitter is a nontrivial 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 [23]. 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 [7], extracting entities [1] or linking tweets to external knowledge categories [22]. However, works show that topic model such as Latent Dirichlet Allocation (LDA) [3] is more effective in identifying topics from short and informal social media language [13]. 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.

3.2.2 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 [20]. 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 [36, 14]. In this work,we just make use of the off-the-shelf work, i.e. SentiStrength [36]. 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.

3.2.3 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 $(d_{m,t}(s_m) = 1.0)$.

4. 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 [26, 30], 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.

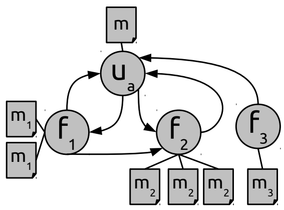

Figure 2: Illustration of relations among tweet, author and followers. Author is denoted as u_a, tweet as m, followers as f_i and tweets of follower f_i as m_i. An directed edge (f_i, u_a) means that f_i is exposed to the messages published by u_a.

4.1 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.

4.2 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.

4.2.1 Topic Similarity

The similarity between two topic distributions can be calculated with methods such as the cosine distance [6] or the Jensen-Shannon Divergence [37]. 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:

$$sim_{topic} = \frac{\theta_m \cdot \theta_u}{\parallel \theta_m \parallel \parallel \theta_u \parallel} \qquad (6)$$

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

4.2.2 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:

$$sim_{opinion}^t(O_t^1, O_t^2) = \frac{8 - |\sum_{i=0}^{8} d_i^1 v_i - \sum_{i=0}^{8} d_i^2 v_i|}{8} \qquad (7)$$

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.

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

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

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.

4.2.3 Subjectivity Similarity

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

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

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.

4.3 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.

5. EXPERIMENTS

5.1 Dataset and Settings

We adopt the Twitter dataset of previous work [21]. 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 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

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

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.

5.2 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) [11] 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	**$4.44e^{-06}$**	0.140	0.272
100	*F*	**43.892**	**31.145**	**28.466**
	p	**$8.65e^{-11}$**	**$3.55e^{-08}$**	**$1.32e^{-09}$**
150	*F*	**22.356**	**12.240**	**14.664**
	p	**$2.43e^{-08}$**	**$6.25e^{-06}$**	**$8.46e^{-07}$**
200	*F*	**31.675**	**20.616**	6.145
	p	**$4.22e^{-06}$**	**$2.92e^{-05}$**	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.

A vivid description about the subjectivity model and its ability in explaining the retweeting behavior can be given with an example. The subjectivity models of a target tweet, its author, and two followers (one retweeter, one non-retweeter) are shown as Figure 3. The tweet talks about topic 14 of the local topic space, and the opinion is neutral. The historical tweets of the author concentrate on the topic 14, and his opinions are mainly neutral. The retweeter has published 195 tweets about two topics (topic 14, 52) and his opinion towards the topic 14 is mainly neutral. While the non-retweeter has also talked about two topics (14th and 56th topic) with 158 tweets, but he is mainly interested in 14th topic, and his opinion is positive. Although the non-retweeter is more similar with both the tweet and author in terms of topic, the retweeter is more similar for subjectivity because his opinion is more approximate with both the tweet and author. The example verifies that the subjectivity model can help better understanding the retweeting behavior by modelling not only topics but also opinions.

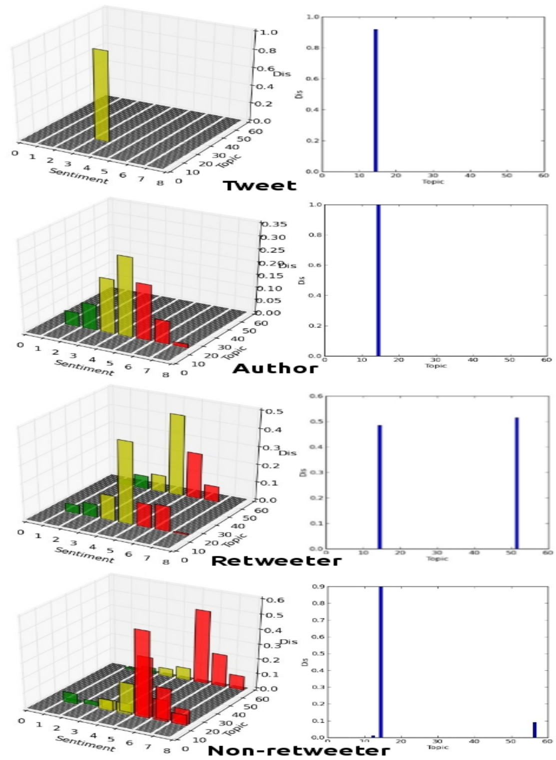

Figure 3: Retweeting anlysis examples.

5.3 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) [1]. 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.* [21] (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 [27] 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.

6. CONCLUSION

Motivated by the psychological research, this paper postulates 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 the users and tweets as well. 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 is put forward in terms of topic simlarity and opinion similarity. The subjectivity model has been 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.

7. REFERENCES

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

[2] A. Asiaee T, M. Tepper, A. Banerjee, and G. Sapiro. If you are happy and you know it... tweet. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1602–1606. ACM, 2012.

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

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

[5] P. H. Calais Guerra, A. Veloso, W. Meira Jr, and V. Almeida. From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158. ACM, 2011.

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

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

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

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

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

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

[12] J. Hannon, M. Bennett, and B. Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *Proc. of the 4th ACM ReSys*, pages 199–206. ACM, 2010.

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

[14] X. Hu, J. Tang, H. Gao, and H. Liu. Unsupervised sentiment analysis with emotional signals. In *Proc. of the 22nd WWW*, pages 607–618. International World Wide Web Conferences

Steering Committee, 2013.

[15] X. Hu, L. Tang, J. Tang, and H. Liu. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 537–546. ACM, 2013.

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

[17] M. Jenders, G. Kasneci, and F. Naumann. Analyzing and predicting viral tweets. In *Proc. of the 22nd WWW*, pages 657–664. International World Wide Web Conferences Steering Committee, 2013.

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

[19] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proc. of the 19th WWW*, pages 591–600. ACM, 2010.

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

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

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

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

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

[25] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd International Web Science Conference*, page 8. ACM, 2011.

[26] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Searching microblogs: coping with sparsity and document quality. In *Proc. of the 20th ACM CIKM*, pages 183–188. ACM, 2011.

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

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

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

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

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

[32] K. Starbird and L. Palen. (how) will the revolution be retweeted?: information diffusion and the 2011 egyptian uprising. In *Proceedings of the acm 2012 conference on computer supported cooperative work*, pages 7–16. ACM, 2012.

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

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

[35] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012.

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

[37] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proc. of the third ACM WSDM*, pages 261–270. ACM, 2010.

[38] Z. Xu, Y. Zhang, Y. Wu, and Q. Yang. Modeling user posting behavior on social media. In *Proc. of the 35th ACM SIGIR*, pages 545–554. ACM, 2012.

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