From Interests to Opinions: Modelling Subjectivity for Retweeting Analysis

Songxian Xie, Jintao Tang and **Ting Wang**
School of Computer, National University of Defense Technology, Changsha, P.R. China

Ruili Wang
School of Engineering and Advanced Technology, Massey University, Auckland, New Zealand

Kewen Wang
School of Information and Communication Technology, Griffith University, Brisbane, QLD 4111 Australia

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. User modelling has been proved to be effective in previous research~~es~~. However few studies have investigated the subjectivity ~~of users~~. Motivated by psychological ~~researches~~, we propose a novel subjectivity model by combining both topic~~ of interests~~ and opinions ~~articulated in UGC~~. We also ~~define a novel weighting function~~ to measure the subjectivity similarity, and demonstrate that a user is more likely to retweet a message because of approximate subjectivity similarity. In the experiments, the subjectivity similarity is verified to be correlated with retweeting behavior by ~~means of~~ ANOVA test. ~~When comparing~~ with other topic-based models in retweeting prediction, our model outperforms ~~other models~~ with best accuracy. ~~Our~~ 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 networking due to the amount of users it has attracted and the volume of messages it produces daily. 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 the subjective users of social networks 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 motivation (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 topics encapsulated in rich linguistic content of a user have been proved consistently dependable for behavior analysis (Petrovic, Osborne, and Lavrenko 2011). However no studies have considered how to model the subjective aspect ("**what's your opinions**") when modelling a user. In this paper, we propose a novel method to model sujectivity of a user(we name it subjectivity model) by combining both the topics and opinions.

Social media is a platform where different opinions are presented by allowing users gennerate contents on topics they are interested in. The subjectivity of a user is encoded in the User-Generated Content (UGC) on Twitter. Therefore, we explore the tweets a user has published to establish the subjectivity model. For the retweeting analysis problem, we assume the probability a user retweets a message could be evaluated ~~with~~ subjectivity similarity measurement. We design a novel weighting function to measure the subjectivity similarity, and propose three subjectivity similarities among a tweet, its author and followers to analyze the motivation of retweeting behaviors. Expertiments on Twitter dataset show that retweeting behaviors are correlated with all three subjectivity similarities, our subjectivity model outperforms topic-based user model for retweeting prediction, and the subjectivity similarities significantly improve the performance of an off-the-shelf predicting model considering other factors.

The rest of the paper is organized as follows: we give the definition and establishment details of the proposed subjectivity model in the next section, and in ~~section~~ 3 the subjectivity similarity is defined and specified for the retweeting analysis problem, experiments of quantitative evaluation is carried out in ~~section~~ 4, the related works are described in ~~section~~ 5 and Section 6 summarizes the paper and points out future work.

Subjectivity Model

Subjectivity has been extensively studied by psychologists to characterize the personality of a person based on his historic 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). How to computationally model the subjectivity of a user is still a challenging problem. 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. First of all, we give a formal definition of the subjectivity model under the context of Twitter.

Definition

Let $G = (V, E)$ denotes a social network, where V is a set of users on Twitter, and $E \subset V \times V$ is a set of following relationships between users. For each user $u \in V$, there is a tweets collection M_u denotes his message history. We assume there is a topic space T containing all topics they talk about, and a sentiment valence space O to evaluate their opinions towards these topics. For the "subjectivity" of a user $u \in V$, we refer to both topics of interest 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 $\{t_i\}$ the user talks about in topic space T and the user's opinions o_i towards each topic evaluated in sentiment valence space O.*

$$P(u) = \{(t_i, w_u(t_i), d_{u,t_i}(o_i)) \,|t_i \in T,\, o_i \in O\} \quad (1)$$

where:

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

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

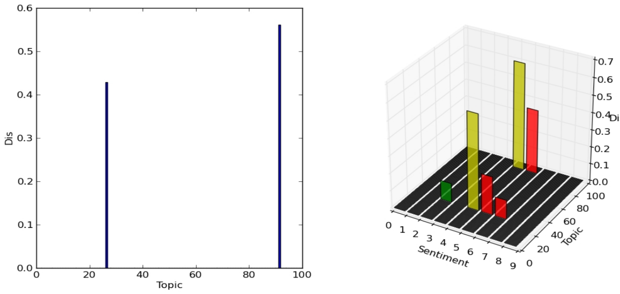

Figure 1: The left subgraph denotes two topics a user tweets about with weights $(w_u(t_{23}) = 0.43, w_u(t_{86}) = 0.57)$. the right subgraph denotes opinions towards two topics: for topic 23, $(d_{u,t_{23}}(o_2) = 0.1, d_{u,t_{23}}(o_4) = 0.5, d_{u,t_{23}}(o_5) = 0.3, d_{u,t_{23}}(o_6) = 0.1)$; for topic 86, $(d_{u,t_{86}}(o_4) = 0.66, d_{u,t_{86}}(o_5) = 0.34)$.

Establishment of Subjectivity Model

The definition of a 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 nontrivial task. There are millions of users and billions of tweets associated with these users. The effectiveness and efficiency of the topic analysis algorithm is a big problem. 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 and network topology on Twitter:

Definition 2 (Local Topic Space) *Let $G = (V, E)$ denotes global social network, 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. In the τ-ego network of u, all users concentrate on limited topics derived from the content generated by them, 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). Works also show that topic models such as **Latent Dirichlet Allocation (LDA)** model (Blei, Ng, and Jordan 2003) are more effective in identifying topics from short and informal social media language (Hong and Davison 2010), so we adopt the user-level LDA model for topic analysis. However, the LDA model is adapted to our local topic space assumption, and the relatively tiny size and topic homophily of an ego network lower the impact of data sparsity and degrade the calculation complexity of LDA.

Opinion Mining In the Natural Language Processing area, 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 (for example: 0 stands for negative and 1 for positive). Sentiment analysis researches have dived into the social media language and provided effective sentiment analysis tools (Thelwall et al. 2010; Hu et al. 2013). In this work,we just make use of the off-the-shelf work of sentiment analysis. The SentiStrength package has been built to cope with sentiment analysis in short informal text of social media (Thelwall et al. 2010). It assigns two values to each tweet standing for sentiment strengths: a positive and a negative sentiment measurement, with $[-5,-1]$ denoting negative and $[1,5]$ denoting positive sentiment strength, which can be used to catch fine opinion distributions in a user's subjectivity model. For the convenience of distribution calculation, we map the output of SentiStrength to a single value in sentiment valence space $[0,8]$ as follows:

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

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

Concreting Subjectivity Model As Definition 2 describes, a τ-ego network $G_u = (U, E_u)$ for a user u can be extracted from global network. Then the subjectivity model of each user $u \in G_u$ can be concreted within the ego network. $M_u = \{m_i | i \in [1, \cdots, N]\}$ denotes tweets set published by user u. A Local Topic Space $T_u = \{t_i | i = 1, \cdots, K\}$ can be constructed using LDA topic model with a single document representing all concatenated tweets in M_u. The topic model is built with the parameter θ_u representing the distribution of user u over topics he talks about. Simultaneously SentiStrength is applied to each tweet $m \in M_u$ and outputs sentiment strength s_m. subjectivity model of user u is built as follows:

- Firstly, the corresponding parameter θ_u of topic model for user u can be regarded as his topic distribution in the Local Topic Space T_u, and the topics he talks about are $Z_u = \{z | p(z|\theta_u(z)) > 0\}$.
- Secondly, the topic model is applied to each tweet m to identify topics it talks about, denoted as $Z_m = \{z_m | p(z_m|\theta, \beta, Z_u) > 0\}$.
- Thirdly, the opinion distribution of user u towards topic $t \in Z_u$ could be calculated as:

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

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

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

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

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.

Retweeting Analysis With Subjectivity Model

Many factors have been proved to affect retweeting behavior (Suh et al. 2010; Macskassy and Michelson 2011; Comarela et al. 2012), however few researches have investigated the subjective motivation of a user to retweet a message. Apart from the context constraints, a tweet is more likely to be retweeted by a user who find its content worth to. Therefore, we are not interested in modelling the tweet by itself as other researches (Naveed et al. 2011; Pfitzner, Garas, and Schweitzer 2012), but the underlying reasons why a user want to disseminate it based on his subjective initiative. We assume that if a tweet is published, all followers of its author will receive it in time, and followers are likely to retweet it if they find it worthwhile. Under such assumption, we investigate the retweeting problem within a 1-ego network for the author of target tweets. In the ego network, the relation among the tweet, its author and followers can be illustrated as Figure 2.

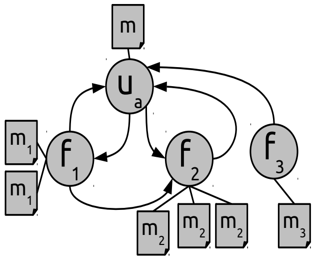

Figure 2: Illustration of relations among tweet, author and followers. Author is denoted as u_a, tweet as m, followers as f_i and tweets of follower f_i as m_i. An directed edge (f_i, u_a) means that f_i is exposed to the messages published by u_a.

Problem Formulation

The retweeting analysis problem can be formulated as follows: For each target tweet m, let F denotes the followers who receive m by following its author u_a, and for each user $u \in F \cup \{u_a\}$, let M_u denotes a tweet collection u has published. For each follower $u_f \in F$, we can define a quadruple $< u_f, u_a, m, r_f >$:

- r_f is a binary label indicating whether m is retweeted by u_f.
- 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_f.

Subjectivity Similarity

We assume that if the tweet content and the subjectivity of a user are similar enough in terms of topics and opinions, the user will have a very high probability to adopt retweeting behavior. With the subjectivity models estabilshed for the author and followers, the subjective decision-making procedure 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 can be divided into topic similarity and opinion similarity. For the calculating convenience, the tweet m is analyzed and a “subjectivity model” is established for m according to the procedure for users. Thus the subjectivity similarity calculation for tweet and user is identical with subjectivity similarity calculation between users.

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) with satisfactory results. We adopt a 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.

Table 1: Illustration of opinion similarity calculation

	0	1	2	3	4	5	6	7	8
$user_1$	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$user_2$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
$user_3$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

Opinion Similarity Opinion in 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, $user_1$ holds most negative opinion on a topic(with 100% sentiment value 0), while $user_2$(100% positive sentiment value 7) and $user_3$(100% positive sentiment value 8) hold positive opinion. If cosine similarity measurement is adopted to calculate opinion similarity, all similarities are 0. In fact $user_2$ is more similar with $user_3$ than $user_1$ because they both hold positive opinion and their sentiment distance is much less than $user_1$ Therefore, opinion similarity can't be calculated simply as the similarity between two distributions. To accurately catch opinion similarity, we propose a novel method by combining both sentiment distance and distribution similarity. The opinion similarity between two users or a tweet and a user on topic t can be calculated as:

$$sim^t_{opinion}(O_1, O_2) = \frac{8 - |\sum_{i=0}^{8} d_i v_i - \sum_{j=0}^{8} d_j v_j|}{8} \tag{7}$$

where d_i denotes the i^{th} entry of opinion distribution vector, and v_i denotes corresponding sentiment strength value. Accordingly, overall opinion similarity on all topics can be calculated as normalized similarity of all opinion similarities on common topics.

$$sim_{opinion}(u_1, u_2) = \frac{\sum_{t=1}^{|T|} sim^t_{opinion}(O_1, O_2)}{|T|} \tag{8}$$

where T denotes the common topics between two users or a tweet and a user, which can be regarded as the intersection between their topic-of-interests.

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

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

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 α, 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 for each user and improve retweeting prediction performance of our model.

Retweeting Analysis

The motivation of retweeting behavior is complicated, which involves the target tweet, its author and followers who is following its author with relations illustrated as Figure 2. The idea behind this work is that taking into account opinions towards interests can yield benefits in explaining the subjective motivation of retweeting behavior. Specifically, given a tweet m, we consider this question from three aspects: (i) how similar is the tweet to the subjectivity of a user in terms of topics and opinions, i.e. $sim_{sub}(m, u)$, (ii) who is the like-minded people for its author among all followers considering similarity of subjectivity, i.e. $sim_{sub}(u_a, u)$, and (iii) whether the tweet is novel or original judged from its similarity with subjectivity of its author, 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 demonstrate such motivation with the three subjectivity similarities.

Experiments

Dataset and Settings

We adopt the Twitter dataset of previous work (Luo et al. 2013), which was created using Twitter API [1]. 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 day. Also each tweet was chosen as starting point to collect at least recent 200 tweets for its author and followers. Overall, there are 4,5531 followers, 6,277,736 tweets, and 5214 followers who have retweeted at least one target tweet during the monitored period. 3,0876 users were extract who have retweeted at least 20 times in their recent tweets to build their subjectivity model and study their retweeting motivations. In order to determine the optimized values of λ for each of the 3,0876 users in subjectivity similarities($sim_{sub}\left(m, u\right)$, $sim_{sub}\left(u_a, u\right)$, $sim_{sub}\left(m, u_a\right)$), we implemented a *mini-batch gradient descent* algorithm with the user's retweeting history. As a result, all three subjectivity similarities are optimized to reflect the personalized retweeting habits of each user.

Correlation of Subjectivity and Retweeting Behavior

The first step of demonstrating our assumption is to assessing the existence of a correlation between subjectivity similarity and retweeting behavior. To verify such correlation, we adopt a statistical hypothesis test called Analysis of Variance (ANOVA) (Fisher et al. 1970). 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. The ANOVA method 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 ANOVA yields a p-value 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 number of topics in LDA as 50, 100, 150 and 200 to determine impact of topic number. The results are listed in Table 2, The bold-faced entries mean that the *p-value* is lower than the target significance level $\alpha = 0.01$.

Note that for the number of topics 100 and 150, all similarities yield *p-values* below α and *F-ratio* above 1.00. This suggests that subjectivity similarities can be good indicators for modeling retweeting behavior. For the rest experiments, we simply set the number of topic as 100 for LDA.

Comparison With Other Models

In this section, we compare our model against other content-based user models such as TF-IDF model(modelling user with bag-of-words) (Luo et al. 2013), entity-based model(model user with entities extracted from the tweets) and hashtag-based model(model user with hashtags used in the tweets) (Abel et al. 2011).

Table 2: ANOVA test for three subjectivity similarities

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

To avoid the bias introduced by imbalance of dataset,a dataset was constructed by taking 5,214 retweeters who retweet at least one target tweet as positive instances, and randomly sampling 5,214 negative instances from the 2,5662 followers who do not retweet any target tweet. The balanced dataset of all positive and negative instances was 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. The accuracy measurement is our evaluation metric, and the result is listed in Table 3.

Table 3: Comparison of Accuracy Performance for Different Models. Significant improvement over hashtag model with $*$, TF-IDF model with † and entity model with ‡ ($p < 0.05$).

Feature	Accuracy(%)
TF-IDF	62.85
entity	68.76
hashtag	59.12
$sim_{sub}\left(m, u\right)$	73.88 $*$ † ‡
$sim_{sub}\left(u_a, u\right)$	70.04 $*$ †
$sim_{sub}\left(m, u_a\right)$	69.64 $*$ †
sim_{all}	**75.64** $*$ † ‡

The best performance (75.64%) is achieved by the sim_{all}, for which We feed all three subjectivity similarities into classifier to test the impact of their combination. The perfromance of TF-IDF model(60.85%) is moderate. Entity-based model (68.76%) is very close to $sim_{sub}\left(u_a, u\right)$ (70.04%) and $sim_{sub}\left(m, u_a\right)$ (69.64%), and their difference is not significant. While for hashtag-based model, its accuracy is the lowest (59.12%), the reason might lie in a very low usage of hashtag in users' tweets. The performance of $sim_{sub}\left(m, u\right)$(73.88%) is better than others except for sim_{all}. The results above show that subjectivity model can better help predicting retweeting behavior than other content-based user models.

Evaluation Considering Other Factors 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 the user, etc. In this section, we firstly compare the performance of our method with

[1] `https://dev.twitter.com/`

model of Luo *et al.* (2013). In their work they use four feature families: “Retweet History”(follower who retweeted a user before is likely to retweet the user again), “Follower Status”(the number of tweets, followers, friends, listed times and verified state), “Follower Active Time”(interaction with other users) and “Follower Interests”(bag-of-words model for users). Then we demonstrate that performance of their method could be improved by using our subjectivity model instead of bag-of-words model in their feature set. We use the same classifier, dataset and settings as last section. The feature set of Luo *et al.* (2013) is marked as “LUO”. In addition,as we note that followers who previously had a history of retweeting might do this in the future. we set a baseline (marked as “RB”), which simply predicts followers who have retweeted the author’s previous tweets as retweeters of target tweet. The result is listed in Table 4.

Table 4: Prediction Accuracy of Different Models. Significant improvement over baseline with $*$ and LUO’ model with ‡ ($p < 0.05$).

Feature Set	Accuracy(%)
RB	60.85
LUO	71.76 $*$
$sim_{sub}\left(m, u\right)$	73.88 $*$ ‡
$sim_{sub}\left(u_a, u\right)$	70.04 $*$
$sim_{sub}\left(m, u_a\right)$	69.64 $*$
sim_{all}	75.64 $*$ ‡
LUO+$sim_{sub}\left(m, u\right)$	74.04 $*$ ‡
LUO+$sim_{sub}\left(u_a, u\right)$	70.27 $*$
LUO+$sim_{sub}\left(m, u_a\right)$	71.86 $*$
LUO+sim_{all}	**78.15** $*$ ‡

The accuracy of baseline is 60.85%, and two prediction methods (“LUO” and our model) both outperform the baseline significantly. $sim_{sub}\left(m, u\right)$ and sim_{all} outperform “LUO” significatantly.

After combining the “LUO” feature set with subjectivity similarities, the accuracy are changing. $sim_{sub}\left(m, u\right)$ gives a significant improvement (LUO+$sim_{sub}\left(m, u\right)$, 2.12% improvement) over “LUO” , which indicates that subjectivity similarity between tweet and followers can be considered as the underlying reason that elicits retweeting behavior. Adding other two subjectivity similarities ($sim_{sub}\left(u_a, u\right)$ and $sim_{sub}\left(m, u_a\right)$) can not improve performance significantly. But noticing that, the most significant improvement(LUO+sim_{all}, 6.39% improvement) is achieved by combining all three subjectivity similarities, which verifies our assumption that a user is more prone to retweet a message with similar subjectivity, like-minded author and original content.

Related Work

User modelling provides insights into user’s online behaviors. Hannon *et al.* (2010) proposed that Twitter users can be modeled by tweets contents and the relation of Twitter social networks, while content-based approach can find similar users who are ”distant” without following relations. Macskassy and Michelson (2011) discovered user’s topic-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.* (2010) made use of topic models to analyze Twitter 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 interest, to explain user posting behavior. Pennacchiotti and Popescu (2011) proposed a most comprehensive method to model Twitter user 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, none of the above works have considered the subjectivity of users, who are subjective initiative individuals in behavior-taking. We have fisrtly proposed a novel subjectivity model for users’ retweeting behavior analysis.

Conclusion

In this paper, from the point of motivation, we postulate that online behaviors of social media users are ~~impacted~~ by their subjectivity. Therefore, we propose a novel model by combining topics and opinions to model the subjectivity of users, which we name as subjectivity model. We design an algorithm for the establishment of subjectivity model, and to make it more efficiently, we consider only the users in an ego network instead of global network and propose a Local Topic Space concept according to the homophily principle. We design novel subjectivity similarity measurement in terms of topic simlarity and opinion similarity. The subjectivity model is applied to the retweeting behavior analysis with three subjecivity similarities among a tweet, its author and followers. Experiment results demonstrate the effectiveness of our model for 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 *Proceedings 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 *Proceedings 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 *Proceedings 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 *Proceedings 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 *Proceedings 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 *Proceedings 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 *Proceedings 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 *Proceedings 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 *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*.

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 *Proceedings 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 *Proceedings 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 *Proceedings 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 *Proceedings of the 19th ACM international conference on Information and knowledge management*, 1633–1636. ACM.