Resonation Elicits Diffusion:
Modelling Subjectivity for Retweeting Analysis

Abstract

Retweeting is the core mechanism of information diffusion on Twitter. User modelling has been proved effective in retweeting behavior analysis, however none studies have investigated the subjectivity of users. In this paper, we propose a subjective model by combining both topics of interest and opinions to model users, and demonstrate that a user is more likely to retweet a message because of subjective similarity. We define a novel weighting function to measure the subjective similarity. By means of ANOVA test, the subjective similarity is verified to be correlated with retweeting behavior; we compare our model with other content-based models in retweeting prediction and results show that our model outperforms other models for predicting retweeting behavior; when combining with other factors, subjective similarities give significant improvement over a off-the-shelf predicting model.

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. Retweeting convention and complex network of Twitter provide an unprecedented mechanism for the spread of information despite the restricted length of a single message(tweet) (Jenders, Kasneci, and Naumann 2013). Actually almost a quarter of the tweets are retweeted from others (Yang et al. 2010). Understanding how retweeting behavior works can help explaining information dissemination on Twitter.

Studies have tried to find factors that influence whether a tweet will be retweeted (Boyd, Golder, and Lotan 2010; Kwak et al. 2010; Suh et al. 2010). 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, estimating the content and deciding to share. The crucial part is to estimate whether a tweet contains information interesting to the user who might find it worthy to be shared. Therefore modelling the motivation of users provides an important perspective for retweeting behavior analysis. This research is motivated by a desire to understand what drives users of social networks to disseminate information they come across.

Previous studies on retweeting behavior 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 ("**Who you tweet**") (Pennacchiotti and Popescu 2011). Subjective initiative nature of human determines that his behavior pattern is subjectivity driven. According to theory of Biased Assimilation, people are prone to choose and disseminate information according to their own biased subjectivity (Hyman 2000). Therefore, we can understand the underlying reason why a user retweet a message by modelling the subjectivity of the user.

Twitter has become a platform where different opinions are presented and exchanged by allowing users publish subjective messages on topics they are interested in. Therefore subjectivity of a user is encoded in the millions of User-Generated Content(UGC) on Twitter. In this study we explore the big text data with state-of-the-art Natural Language Processing techniques to model the subjectivity of users, and investigate whether the subjective model could benefit the retweeting behavior analysis. Intuitively, we put forwards that subjectivity can be represented as topics and opinions articulated in the tweets generated by users on Twitter. And the probabilty a user retweets a message could be evaluated by measuring the subjective similarity between the tweet and the subjective model of the user.

Subjective 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 languistic content (Stein and Wright 2005). How to computationally model subjectivity of a user is still a challenging problem for many applications. The advent of online social networking such as Twitter has given a new layout to the challenge. Twitter allows users to show their personal subjectivity by publishing short messages, which give researchers resources to model the subjectivity of users. First of all, we give an formal definition for subjective model under 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 directed 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 for evaluating their opinions towards these topics. For the "subjectivity" of a user u, we refer to both topics of interest and opinions articulated in his tweets collection M_u.

Definition 1 (Subjective Model) *The subjective model $P(u)$ of a user $u \in V$, is the combination of a set of topics $\{t_i\,(i \in \{1 \cdots n\})\}$ the user talks about in a topic space T and the user's opinion o_i towards each topic t_i.*

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

where:

- *with respect to the given 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, with $\sum_{i=1}^{|T|} w_u(t_i) = 1$.*
- *opinion of the user towards topic t_i is modeled as a topic-dependent sentiment distribution $d_{u,t_i}(o_i)$ over sentiment valence space O.*

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

Establishment of Subjective Model

According to the definition, subjective model of a user can be represented as two distributions: the topic distribution and the sentiment distribution for each topic. With state-of-the-art topic model and sentiment analysis techniques, we can establish subjective model by finding topics and opinions simultaneously.

Topic Analysis Previous studies have tried to identify topics 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). Recent works show that topic models such as **Latent Dirichlet Allocation (LDA)** model (Blei, Ng, and Jordan 2003) have been efficient ways to characterize latent topics of large volume corpus. Topics of LDA are broader in concept, since a single topic consists of the whole collection of related words, so we adopt the LDA topic model for 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 assocated with these users, and the effectiveness and efficiency of topic model is a big problem. However, the "following" 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 topics of interest. According to the principle of homophily, we put forwards the concept of **Local Topic Space**:

Definition 2 (Local Topic Space) *Let $G = (V, E)$ denotes a 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 a 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 user u, all users concentrate on a few topics derived from the content generated by them, and these topics form a local topic space T_u.*

The relatively tiny size and topic homophily of ego network lower the impact of data sparsity and degrade the complexity of calculation.

Therefore we adopt a user-level LDA model to find latent topics in Local Topic Space by aggregating all tweets generated by a user into a single document just as (Hong and Davison 2010). The result of LDA produces two multinomial distributions, with distribution θ_u representing the probability distribution of a user over each topic, and distribution β_k representing the probability distribution of a topic over whole vocabulary used by all users in ego network.

Opinion Analysis Users often express opinions towards topics of interest by publishing topic-related tweets. Sentiment analysis or opinion mining is formally defined as the computational study of sentiments and opinions about topics expressed in a text (Liu 2012). Opinions are regulated as sequential discrete values to represent sentiment strength(for example: 0 stands for negtive and 1 for positive). Recently, researchers began to pay more and more attention to social media such as Twitter (Thelwall et al. 2010; Thelwall, Buckley, and Paltoglou 2012; Hu et al. 2013) and provide effective sentiment analysis tools.

SentiStrength package has been built especially to cope with sentiment analysis in short informal text of social media (Thelwall et al. 2010). It combines lexicon-based approaches with sophisticated linguistic rules adapted to social media, which is suitable for analyzing sentiment of tweets. SentiStrength 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. Sentiment assigned by SentiStrength is a fine-grained strength, which can catch fine opinion distributions in a user's subjective model. For the convenience of distribution calculation, we map the output of SentiStrength to single-scaled sentiment valence space $[0, 8]$ as follows:

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

Where p denotes the positive setiment strength and n denotes negative sentment strength. In the sentiment valence space, value 4 indicates neutral sentiment, while values above 4 indicate positive sentiment and values below 4 indicate negative sentiment.

Concrete Subjective Model As Definition 2 describes, for a user u in a global netwotk, we can extract a τ-ego network $G_u = (U, E_u)$. And the subjective model of each user $u \in G_u$ can be concreted within the ego network. We denote tweets set published by a user u as $M_u = \{m_i | i \in [1, \cdots, N]\}$. All tweets in M_u is concatenated to a single document, and to construct Local Topic Space $T_u = \{t_i | i = 1, \cdots, K\}$ with LDA topic model. The topic model is built with parameter θ_u representing the distribution of user u over topics he talks about, and parameter β_k represents the distribution of each topic over the vocabulary of all tweets. SentiStrength is applied to each tweet m in collection M_u and outputs sentiment strength s_m for tweet m. Subjective model of user u is built the as follows:

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

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

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:

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

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

where s_m denotes the sentiment strength of tweet m. For simplicity, we assume the sentiment of tweet m is related to every topic it talks about in Z_m.

Totally, we build subjective model $P(u)$ for user u as:

$$P(u) = \{(t, p(z|\theta_u), d_{u,t}(o)) | t \in Z_u, o \in O\} \qquad (6)$$

Retweeting Analysis With Subjective Model

Many factors have been proved to influence 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. Therefore we will study whether subjective model can help understand underlying reasons of a user's retweeting behavior. Apart from the context constraints, a tweet is more likely to be retweeted by subjective users who find it worth to. Therefore, we are not interested in modelling the tweet by itself as other researches (Naveed et al. 2011b; 2011a; Pfitzner, Garas, and Schweitzer 2012), but how the tweet resonate with the users who might want to disseminate it. We assume the motivation of a user to retweet a message lies in that the user considers only the tweet content arousing his resonation without context perturbation. 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 problem in 1-ego network for the author of target tweet. The relation between tweet, author and followers can be illustrated as Figure 1.

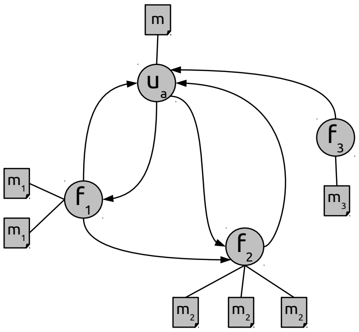

Figure 1: Illustration of relations between tweet, author and followers. Author is denoted as u_a, tweet as m, followers as f_i and tweets of follower f_i as m_i. An directed edge (f_i, u_a) means that f_i is exposed to the messages published by u_a.

Problem Formulation

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 tweets the user 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 subjective 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 subjective similarities between tweet m, its author u_a and follower u_f.

Subjective Similarity

In order o understand the underlying reasons why a user retweet a message, we try to simulate the subjective decision-making procedure by investigating the relationship among a tweets and subjective models of its author and followers. We assume that if a tweet and the subjective models of a user are similar enough in terms of topics and opinions, the user will have a very high probability to retweet the tweet. We call this phenomenon as "resonation", and assume resonation between tweets and users will elicit retweeting behavior. With subjective models built for users, we can define a similarity measurement to quantify the resonation among them.

Topic Similarity Topic distribution of a tweet can be inferenced by applying topic model estimated from the tweets collections of users. The topic similarity between tweet and user can be calculated with methods such as the cosine distance (Cha 2007) or the Jensen-Shannon Divergence (Weng et al. 2010) with satisfactory results. In our research, we adopt cosine distance to measure the topic similarity, which is defined as:

$$sim_{topic} = \frac{\theta_m \cdot \theta_u}{\parallel \theta_m \parallel \parallel \theta_u \parallel} \tag{7}$$

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

Opinion Similarity Opinions of a tweet can be analyzed with SenStrength and the values are transformed to range $[0, 8]$ with Equation 2. Opinions of a tweet towards each topic it talks about can be regarded as a distribution with 1.0 probability on a single sentiment value just as opinion dsitribution of subjective model. Therefore opinion similarity between tweets and users is the same with similarity between two users. We treate opinion as distribution over sentiment valence space with each element of the distribution represents the propotion of the corresponding strength values in the user's all sentiment values. However ,values in sentiment valence space are not independent. They are sequential and represent strength of the sentiment, for example value 8 represents more positive sentiment than value 5, and their sentiment distance is 3. Therefore, opinion similarity can not be calculated as the distance between two distributions. 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 $user_1$. In order to accurately catch opinion similarity, we propose a novel similarity-calculating method by combining both sentiment distance and distribution similarity. The opinion similarity of two users or a tweet and a user on topic t can be calculated as:

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

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

where d_i denotes the ith distribution of opinion 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 their common topics of interest.

$$sim_{opinion}(u_1, u_2) = \frac{\sum_{t=1}^{|T|} sim^t_{opinion}(O_1, O_2)}{|T|} \tag{9}$$

where T denotes the common topics of interest between two users or a tweet and a user.

Subjective Similarity Given a tweet t and a user subjective model $P(u)$, we can define their subjective similarity by combining topic similarity and opinion similarity as follows:

$$Sim_{sub}(t, u) = \alpha * sim_{topic} + (1 - \alpha) * sim_{opinion} \tag{10}$$

where α is the coefficient used to control the proportions of topic similarity and opinion similarity in the holistic subjective similarity. A user cares more about topics a tweet talks about with a larger α, and cares more about opinions with a samller α. A personalized α can be learned from the retweeting history of a user to represent his retweeting habit.

Retweeting Analysis

The motivation why a user retweet is complicated, and involves the tweet itself, its author and its receivers who is following its author. Figure 1 illustrates their relation. The idea behind this work is that taking into account user attitudes towards his own interests can yield benefits in explaining the subjective motivation of retweeting behavior. Specifically, we consider this question from three aspects: (i) how similar is the tweet to the subjective model of a user in terms of topics and opinions, i.e. $sim_{sub}(m, u)$, (ii) are the author and user like-minded people considering similarity of their subjective models, i.e. $sim_{sub}(u_a, u)$, and (iii) whether the tweet is novel or original judged from its similarity with subjective model of its author, i.e. $sim_{sub}(m, u_a)$. Intuitionally, a user might retweet a message if its content is close enough to his subjectivity, its author is a like-minded friend and it is original from inner subjectivity of its author. In this work we will carry out a sery of experiments to demonstrate such intuition.

Experiment

Dataset

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 few days. And 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. We extract 3,0876 users who have retweeted at least 20 times in their recent tweets to build their sujective model and study their retweeting motivations. In order to determine the optimized values of

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

α in subjective 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 users' retweeting history. As a result, all three subjective similarities are optimized to reflect the personalized retweeting habits of each users.

Correlation of Subjectivity and Retweeting Behavior

The first step of demonstrating our intuition is to assessing the existence of a correlation between subjective similarity and retweeting behaviour. To test this correlation, we adopt a statistical hypothesis test called Analysis of Variance (ANOVA). 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 if the distinct sets of retweeters and non-retweeters do have the same subjective similarity to a tweet that both sets are exposed to. 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. If ANOVA yields a p-value lower than the significance level α, the *null hypothesis* is rejected. 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,

Table 2: ANOVA test for three subjective 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

Note that for the number of topics 100 and 150, all similarities yield *p-values* below α. This suggests that subjective similarities can be good indicators for modeling retweeting behavior. Therefore we reject the *null hypothesis* and accept the alternative hypothesis. For the rest experiments, we simply set the number of topic as 100 for LDA.

Comparison With Other Models

Given that ANOVA indicated correlation between subjective similarities and retweeting behavior, we compare three similarities against other content-based similarity measurements such as TF-IDF model(modelling user with bag-of-words in his tweets) (Luo et al. 2013), entity-based model(model user with entities extracted from his tweets) and hashtag-based model(model user with hashtags used in his tweets) (Abel et al. 2011).

We take 5,214 retweeters in our dataset who retweet at least one target tweet as positive instances, and randomly sample 5,214 negative instances from the 40,317 followers who do not retweet any target tweet. The balance dataset of all positive and negative instances is randomly divided into five parts for 5-fold cross-validation. We use the logistic regression classifier of Scikit-learn machine learning package (Pedregosa et al. 2011) for training and testing. Accuracy is our evaluation metric, and the result is listed in Table 3.

Table 3: Comparison of prediction accuracy for different models.

Feature	Accuracy(%)
TF-IDF	60.85
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 accuracy(75.64%) is achieved by the sim_{all}, for which We put all three subjective similarities into classifier to test 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 subjective model can better help understand retweeting behavior than other user models.

Combining Evaluation With Other Factors Subjective model tries to catch the subjective motivation of users based on their tweets content, 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 combine subjective similarities with other important factors that other researches have demonstrated. Firstly we compare the performance of our method with 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 being verified), "Follower Active Time"(interaction with others) and "Follower Interests"(bag-of-words used to model users). Secondly we demonstrate that performance of retweeting prediction could be improved significantly by using subjective model instead of bag-of-words model in their feature set. We use the same logistic regression classifier, dataset and settings as last experiment. The feature set of Luo *et al.* (2013) is marked as "LUO". In addition, 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.

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.

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

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

After combining subjective similarities, the accuracy are changing. Noticing that, the most significant improvement(LUO+sim_{all}, 6.39% improvement) is achieved by adding all three subjectivie similarities, which verifies our intuition. Subjective similarity between tweet and followers also gives significant improvement(LUO+$sim_{sub}\,(m, u)$, 2.12% improvement), which indicates that subjective resonation between tweet and followers can be considered as the underlying reason that elicits retweeting behavior. Adding other two subjective similarities can not improve performance significantly.

Case Study

In this section we give an vivid description about subjective model and its ability in explaining the retweeting behavior with an example. Topic and opinion of one target tweet, subjective models for its author, and two followers (one retweet it while the other does not) are shown as Figure 2. The right part of each sub-figure illustrates topic distribution and the left part illustrates opinion distribution. It is clear that the tweet talks about the 14th topic of the local topic space. Figure 3 shows top words of the 14th topic, the tweets of author and followers with word cloud[2]. Content of the tweet is:

Tweet: "Sometimes the right person for you was there all along. You just didnt see it because the wrong one was blocking the sight"

The topic of this tweet is about "love between people" and the opinion is neutral, which is in accordance with the 14th topic word cloud in Figure 3 and sub-figure of tweet in Firgure 2. The author concentrates on the 14th topic, and his opinions are mainly neutral (as Figure 2, 3 show). As for two followers, the "retweeter", has published tweets about two topics (the 14th and 52nd topic) uniformly and his opinions towards the two topics are mainly neutral. While the "unretweeter", has also talked about two topics (14th and 56th topic), but he is mainly interested in 14th topic and has positive opinion. Although two followers have same interest (the 14th topic), their different opinions elicit their different decision, which verifies subjective model can help better understand the retweeting motivation by modelling not only topics but opinions.

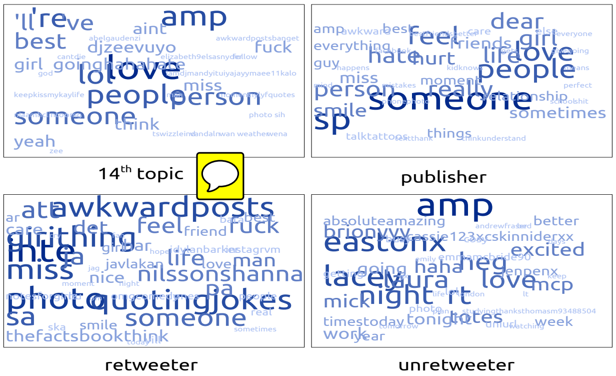

Figure 3: Word cloud of 14th topic, publisher and followers

Related Work

User modelling provides researchers with insights into user's online behaviors. Hannon *et al.* (2010) proposed that Twitter users can be modeled by tweets content and the relation of Twitter social network, while content-based approach could find similar users who are "distant" without follow relations. Macskassy and Michelson (2011) 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.* (2010) made use of topic models to analyze Twitter content 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; Starbird and Palen 2012), 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).

Conclusion

In this paper, we propose subjective model to model users, and demostrate its ability in retweeting behavior analysis. We assume that retweeting should be elicited by the subjective similarities among the tweet and its author and followers. We define subjective model formally as the combination of topic distribution and opinion distribution, and we concrete subjective model leveraging statistical topic model

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

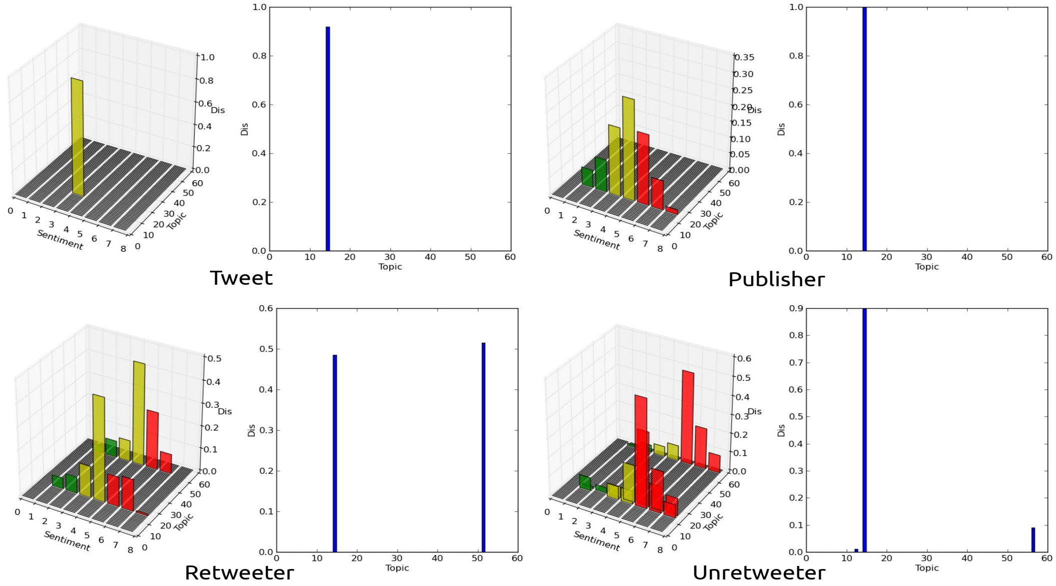

Figure 2: Subjective model examples.

and sentiment analysis techniques. We demonstrate the effectiveness of our model for retweeting analysis problem and show that subjective model is able to reach better understanding of retweeting behavior.

In the future, we will apply subjective 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 *Proceedings of the 19th international conference on User modeling, adaption, and personalization*, UMAP'11, 1–12. Berlin, Heidelberg: Springer-Verlag.

Blei, D.; Ng, A.; and Jordan, M. 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 *2010 43rd Hawaii International Conference on System Sciences*, volume 0, 1–10.

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*, CHI '10, 1185–1194. New York, NY, USA: 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*, HT '12, 123–132. New York, NY, USA: 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):1261–1268.

Feng, W., and Wang, J. 2013. Retweet or not?: personalized tweet re-ranking. In Leonardi, S.; Panconesi, A.; Ferragina, P.; and Gionis, A., eds., *WSDM*, 577–586. ACM.

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*, RecSys '10, 199–206. New York, NY, USA: 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*, SOMA '10, 80–88. New York, NY, USA: 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*, WWW '13, 607–618. Republic and Canton of Geneva, Switzerland: 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 compan-*

ion, WWW '13 Companion, 657–664. Republic and Canton of Geneva, Switzerland: 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*, WWW '10, 591–600. New York, NY, USA: 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 Adamic, L. A.; Baeza-Yates, R. A.; and Counts, S., eds., *ICWSM*. The AAAI Press.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27(1):415–444.

Naveed, N.; Gottron, T.; Kunegis, J.; and Alhadi, A. C. 2011a. Bad news travel fast: A content-based analysis of interestingness on twitter. In *WebSci '11: Proceedings of the 3rd International Conference on Web Science*.

Naveed, N.; Gottron, T.; Kunegis, J.; and Alhadi, A. C. 2011b. Searching microblogs: coping with sparsity and document quality. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, 183–188. New York, NY, USA: 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 *International AAAI Conference on Weblogs and Social Media*.

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 Breslin, J. G.; Ellison, N. B.; Shanahan, J. G.; and Tufekci, Z., eds., *ICWSM*. The AAAI Press.

Ramage, D.; Dumais, S.; and Liebling, D. 2010. Characterizing microblogs with topic models. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. AAAI.

Starbird, K., and Palen, L. 2012. (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*, CSCW '12, 7–16. New York, NY, USA: ACM.

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 *Proceedings of the IEEE Second International Conference on Social Computing (SocialCom)*, 177–184. Minneapolis: IEEE.

Thelwall, M.; Buckley, K.; Paltoglou, G.; Cai, D.; and Kappas, A. 2010. Sentiment in short strength detection informal text. *J. Am. Soc. Inf. Sci. Technol.* 61(12):2544–2558.

Thelwall, M.; Buckley, K.; and Paltoglou, G. 2012. Sentiment strength detection for the social web. *J. Am. Soc. Inf. Sci. Technol.* 63(1):163–173.

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*, SIGIR '12, 545–554. New York, NY, USA: ACM.

Yang, Z.; Guo, J.; Cai, K.; Tang, J.; Li, J.; 0007, L. Z.; and Su, Z. 2010. Understanding retweeting behaviors in social networks. In Huang, J.; Koudas, N.; Jones, G. J. F.; Wu, X.; Collins-Thompson, K.; and An, A., eds., *CIKM*, 1633–1636. ACM.