Resonation Elicits Retweet: Modeling Subjectivity for Retweet Analysis

ABSTRACT

Retweet is the core mechanism of information diffusion on Twitter, and many factors have been proved to influence rewteet behavior, however few studies have investigated the subjective motivation of a user to retweet a message. In this paper, in the light of psychological theory, we assume that a tweet is retweeted by a user because of similar subjectivity, and propose a subjective model to combine both the topics of interest and opinions of users. Technically, We use the state-of-the-art Latent Dirichlet Allocation topic model to find the topics users are talking about, and sentiment analysis techniques to determine user's opinions towards these topics from user-generated content simultaneously. We evaluate our model in the retweet analysis problem to verify its influence on retweet behavior and effectiveness in the retweet prediction performance.

1. INTRODUCTION

Twitter is well-known for its freedom of publishing short message(i.e. tweet) within a limited length of 140 characters, and viral spreading of information across complex social networks. Since launched in March 2006, the service rapidly has gained worldwide popularity, with over 500 million registered users at the end of year 2012, who have generated over 340 million tweets per day[1]. Beyond the large collection of user-generated content, Twitter provides its social network functions for connection, communication, and information diffusion by allowing users to message one another

---

[1]http://en.wikipedia.org/wiki/Twitter

directly and follow one another publicly. The complex networks and large content volume of Twitter provide researchers with insights into people's social behavior on a scale that has never been possible[31].

Information diffusion is a lasting research for social scientists who might want to study the problem on Twitter because an established convention of Twitter facilitates information diffusion. Despite the restricted length of tweets, retweet convention and complex networks of Twitter provide an unprecedented mechanism for the spread of information[14]. In fact almost a quarter of the tweets published by users are retweeted from others[38]. Therefore, it is important to make it clear how retweet works in order to understand how information is diffused on Twitter. This paper focuses specifically on how information spreads on Twitter by exploring underlying reasons why a user retweet a particular tweet out of the overwhelming message flow.

Several studies on retweet have concentrated on analyzing retweet habits and influencing factors[4, 16, 32], most of which are generic, not user oriented. From the point of users, the practice of retweet includes reading the tweet, deciding whether it is worthwhile to share, and then acting on it, hence retweet behavior indicates that the user considers the tweet containing information interesting to him. Therefore modeling users provides an important perspective for retweet behavior analysis. Studies on the retweet behaviors of users have shown that an enriched user model gives coherent and consistent explanation for retweet behavior[1, 18, 9]. Until now, researchers have tried to model users from four types of information to characterize a Twitter user: profile features("**Who you are**"), tweeting behavior("**How you tweet**"), linguistic content("**What you tweet**"), and social network("**Who you tweet**")[25]. Despite demographic profile, tweeting habits and network connections might determine source and scope of information users could be exposed to, topics of interest of users encapsulated in rich linguistic content have been proved consistently valuable for retweet bahavior explanation[18].

Beyond merely publishing news and reports, Twitter is becoming a large platform where different opinions are presented and exchanged by allowing users comment, discuss, compliment, argue and complain over topics they are interested in freely. Studies have demonstrated that user-generated content with rich sentiment information can trigger more attention, feedback or participation[31], and tweets with high emotional diversity have a better chance of being retweeted[26]. However, users receive thousands of tweets about different topics every day, which one will be retweeted depends on the subjective choice of users. Philosophically, subjective initiative nature of human determines that the pattern of behavior is subjectivity driven. Psychological researchers have identified subjectivity as the underlying factor that influence the decision-making

about taking what activities to process incoming stimuli[20]. According to theory of Biased Assimilation, bias is an inherent feature of human subjectivity, and bias is a consistent, robust characteristic of behavior, so people are prone to choose and diffuse information according to their own biased subjectivity[13, 33]. It can be inferred from above explanation that users are prone to diffuse a tweet that can raise resonation with their biased subjectivities. On Twitter, subjectivity may be represented as topics and opinions articulated in the information generated by users. Inspired by this observation, we explore the linguistic information of Twitter to model the subjective tweets and users, and investigate how the subjective model could benefit the retweet behavior analysis.

Establishing subjective model on Twitter faces a number of challenges, including sparsity of information, dynamics of topics and opinions, informal nature of language, etc. However, we are interested in understanding retweet behavior at a local level rather than at a global level as most of time retweet behavior pertains to a local network of the tweet publisher and followers, thus the relatively tiny size of local network and topic homophily lower the impact of sparsity. Given the biased nature of subjectivity, while new information may arise and old information may change their meaning, biased subjectivity to be more consistent and less prone to external perturbations, unless users actually change their interests and opinions, which is usually a relatively long process. Our work aims to define and establish the subjective model and identify the role of subjectivity in the processes of information diffusion on Twitter. By using "**Subjectivity**" , we refer to both topics of users' interests and users' opinions towards these topics, so we model users not only by topics they care about, but also by "**what they think about the topics**". Our contributions can be summarized as follows:

- Based on psychological theory, we put forward formal definition of subjective model for users and tweets.
- Based on state-of-the-art topic model and sentiment analysis technique, we propose the method of establishing subjective model from user-generated content on Twitter.
- Combined with the retweet behavior analysis problem, we analyze the impact of subjective model on retweet behavior and compare our model with other modeling methods.

The rest of the paper is organized as follows: Section 2 gives related work to our research, the proposed subjective model is defined and specified in section 3,the qualitative and quantitative evaluation is described in section 4. Section 5 summarizes the paper and points to future work.

2. RELATED WORK

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

2.1 Retweet behavior analysis

A large body of studies have analyzed characteristics of of retweet, examining factors that lead to increased retweetability and designing models to predict the probability of being retweeted. As for factors influencing retweetability, [32] 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. [18] studied a set of Twitter users over a period of a month and sought to explain the individual rewteet behaviors. They found that models derived from tweets content could explain most of retweet behaviors. [7] found previous response to the tweeter, the tweeters' sending rate, the freshness of information, the length of tweet could affect followers' response to retweet. [29] addressed specifically the retweet mechanism during crises. They found that tweets with topical keywords were more likely to be retweeted. There are also works extending the analysis to build retweet prediction model. [38] designed features from factors that influence the retweet behavior and built a semi-supervised framework on a factor graph model to predict users' retweet behaviors. [23] 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. They found the dominance of social features, while tweet features added a substantial boost to the performance. [14] analyzed the "obvious" and "latent" features from structural, content-based, and sentimental aspects of both tweet 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. [22, 21] 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 . [9] 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 re-rank the tweets according to their probability of being retweeted. [26] 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 listed above have tried to answer the question of "Which tweets will be retweeted by anyone or not?". 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 is capable of capturing both the interests and opinions of users.

2.2 User profile and user modeling

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, which provide researchers with insights into user online behaviors. [10] proposed that Twitter users can be modeled by the tweets and relationships of Twitter social network. They found that social graph based approach could explore users who are 'near' in the sense of follow relations, but failed to find users who are 'distant', on the other hands, content-based approach could find similar users based on interests extracted from the content of tweets. [18] 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. [27] made use of topic models to analyze Twitter content at large scale and at the level of individual users with 4S dimensions, showing improved performance on tasks such as post filtering and user recommendation. These efforts about user modeling on Twitter have simply built model for each user by extracting key words, 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. [37] proposed a mixture model which incorporated three important factors, namely breaking news, friends' time-line and user interest, to explain user posting behavior. [25] proposed the most comprehensive method to model Twitter user for user classification. They focused on richer feature sets such as features derived from topic models,

tweet sentiment analysis and explicit follower-followed links to exploit the user-created content. Their work confirmed the value of in-depth features which reflect a deeper understanding of the Twitter user and the user network structure.

Just as the Section 1 stated, previous researches have tried to model users from four types of information: profile features, tweeting behavior, linguistic content and social network. Some studies perceive that the implicit features articulated in the user-generated content play an import role in user behavior analysises, and they have proposed diverse techniques to capture such in-depth features to model user's interest. Additionally, a few works even identify the correlation between sentiment of users and their behaviors. But they all fail to model subjectivity of a user as a whole. Motavited by the observation, we firstly put forward subjective model to combine both interests and opinions to model a holistic user.

2.3 Sentiment Analysis

Sentiment analysis has been a popular researching area for years, and previous research mainly focuses on reviews or news comments. Generally, there are two major methods, i.e. rule-based approaches and machine learning approaches. Recently, the sentiment analysis research has begun to pay more and more attention to social media such as Twitter because of the massive user-generated content and the unique characteristics which could be incoporated into sentiment analysis. [11] 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. [15] focused on target-dependent Twitter sentiment classification, they proposed to improve target-dependent Twitter sentiment classification by taking target-dependent features and related tweets into consideration. [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. [12] 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, [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. [35, 34] designed SentiStrength, an algorithm for extracting sentiment strength from informal English text with human-evaluated dictionaries for words connotated with positive or negative sentiment strength, which exploited the grammar and spelling styles in typical microblogs. 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. SUBJECTIVE MODEL

In this section, we firstly give definition of subjective model and describe the techniques of building subjective model elaborately, then formulate the retweet analysis problem with subjective model.

3.1 Definition

Psychologically, the subjectivity of human is the underlying reason that drives his social behavior. When a user browses a tweet in the message flow he has followed, the inner subjectivity defines his retweet choice for the tweet, which depends on whether the the topics and the opinions of the tweet could arouse resonation of the user. If the tweet is similar with the user in both topics and opinions, the user might have high probability to propagate it to his followers by taking retweet behavior. The core problem lies in that how to model the subjectivity of the user and the tweet(we refer subjectivity of tweet as the topics it talks about and the opinions towards these topics it holds). Subjectivity has been extensively studied by psychologists from different dimensions to characterize the personality of people based on such records as historic behaviors and remarks[8], and linguists have defined the subjectivity of language as the speaker always shows their perspectives, attitudes and sentiments in the discourse[30]. The user-generated content of social media provide massive language data that traditional collection methods can not achieve. With plenty of data available, we firstly propose and define subjective model on Twitter, and we declare that our definition can be easily be adapted for other social media platforms as well.

DEFINITION 1 (SUBJECTIVE MODEL FOR USER). *The subjective model of a user* $u \in U$ *is a set of topics* t_i ($i \in \{1 \cdots n\}$) *the user talks about in a topic space* T *and the user's opinions* o_i *towards these topics.*

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

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.*
- *opinion* o_i *of user towards topic* t_i *is a target-dependent sentiment distribution* $d_{u,t_i}(o_i)$ *over sentiment valence space* O.

As medium for the users to express themselves, tweets are also subjective to some extent. In order to quantitatively measure the degree of similarity betweet a tweet and a user, the tweet should also be modeled in the same form as the user. Accordingly, the subjective model of a tweet is define as following:

DEFINITION 2 (SUBJECTIVE MODEL FOR TWEET). *The subjective model of a tweet* $c \in C$ *is a set of topics* t_i ($i \in \{1 \cdots n\}$) *it talks about in the same topic space* T *as the users, and the opinions* o_i *it expresses towards these topics.*

$$P(c) = \{(t_i, w_c(t_i), d_{c,t_i}(o_i)) \mid t_i \in T, o_i \in O\} \quad (2)$$

3.2 Retweet Problem Statement

Although subjective model might explain several behaviors of social media users, we are interested in retweet behavior in this paper, and apply subjective model to understanding underlying mechanism of information diffusion at a micro-level. It has been demonstrated that any retweeted tweet is able to reach an average of 1,000 users no matter how many followers the original tweet has[16]. Retweet function has given every user the power of spreading information broadly, and individual user has the power to dictate which information is important and should spread by pushing retweet button. In fact a user only retweets a small number of tweets and only small subset of followers actually retweet a tweet because the likelihood of a tweet to be retweeted depends on both network context and tweet content. Apart from network context, a tweet is more likely to be retweeted by users are who are interested in the tweet, therefore, we are not only interested in modeling the tweet by itself

but also how the tweet resonates with the individuals who might decide to pass it on, and we put a much stronger emphasis on the content and try to model the user's retweeting decision by deriving high-level content-based topic and opinion features.

In fact whether a tweet is retweeted depends heavily on context such as the author's position in the social graph or the time of day the tweet is published. A tweet with only few or passive followers is less likely to be retweeted, and tweets published in the night have less chance to get retweeted than daytime. Despite such a fact, neither contextual factor has any influence on the content of a tweet, therefore we deliberately ignore context information to avoid introducing contextual bias into our analysis of rewteet by proposing a hypothesis.

HYPOTHESIS 1 (H1). *A tweet is evenly visible to the followers who subscribe to it by following its publisher.*

The rationale behind this hypothesis is, that the motivation of a user to retweet a tweet is that the user considers only the tweet content that arouse resonation with the user.

Studies have highlighted that Twitter network structure better resembles an information sharing network than a interconnecing social network[4, 16]. In the context of Twitter, the "following" relationship is a strong indicator of a phenomenon called "homophily", which has been observed in many social networks. Homophily is a phenomenon that people of social network "are homogeneous with regard to many socio-demographic, behavioral, and intra-personal characteristics"[19]. In other words, homophily implies that a user follows another user because he is interested in what another user talks about in tweets, and another user follows back because he finds they share similar interests. According to the principle of homophily, we put forwards the concept of **Local Topic Space** that the content generated by users in the local network of a user and his followers concentrate on a few local topics.

If a tweet is published, all followers will receive it in their timeline. We suppose every follower read it and have the chance to retweet it if they find it worth to according to our hypothesis 1. Ignoring the influence of context information, the goal of our study is to make it clear whether the subjectivity of tweets and users influence retweet, and if so, the analysis of retweet will be reinforced by taking the subjective factor into account. The problem we study can be stated as follows:

Let F, P, C denote the follower set, publisher set and tweet set respectively. As a tweet is created by only one publisher but can be received by multiple followers, each tweet can be defined as a tuple $< f, p, c, r_{fpc} >$ where:

- $c \in C$ is a tweet, $p \in P$ is the publisher of the tweet, and $f \in F$ is a follower of the tweet.
- r_{fpc} is a binary label indicating whether the tweet c is retweeted by a follower f.

Our work focuses on the applicability of content-based subjective model as a static quality measure to analyze the relation between the subjectivity of a user and his retweet beavior. So we could project the tuple into the Local Topic Space T, which is determined with the historic data published by the publisher p and the followers $\{f | f \in F\}$, and represent f, p, c with their subjective models established in T to analyze their relations with the retweet label r_{fpc}.

3.3 Retweet Problem Formulation With Subjective Model

According to definition of subjective model, there are two distributions to model the subjectivity: one is topic distribution, and the other is opinion distribution for each topic, both of which need to be inferred from historic data produced by users. As an open platform for collecting data, Twitter has become a valuable source for quantitative researchers in the last few years. However, content analysis on Twitter poses unique challenges: tweets usually are very short with limit of 140 characters, informal languages are widely used unlike the standard written English, which make many supervised models of machine learning and Natural Language Processing invalid. Effectively modeling content on Twitter requires techniques that can readily adapt to these challenges and require little supervision. In this section, we firstly proceed with detailed methods of how to establish subjective model, and then describe how to apply subjective model to the analysis of retweet behavior.

3.3.1 Local Topic Space

As stated above, the connection relation between publisher and followers indicates their common interests in a Local Topic Space. However, topics usually are not explicitly expressed by users, the topics of a tweet are latent features and have to be inferred by analyzing its content. The distillation of topics from tweet content requires methods that are suitable for short texts with heterogeneous vocabulary. Previous efforts have tried to infer topics by finding key words[6], extracting entities[1], and linking tweets to external knowledge categories[18]. However, the sparsity is a main problem for these methods to model the users' interests because even users have common local topics they still might refer to a topic with different expressing way. Recent works show that topic models such as **Latent Dirichlet Allocation (LDA)** model and its extensions[3, 36] have been efficient ways to characterize latent topics of large volum corpus. Topics of topic models are broader in concept than individual word, since a single topic consists of the whole collection of related words. Thus, the latent topics are more suitable to understand the common topics of interest among different users who use various words to describe the same topic. Therefore we use LDA model to find latent topics that are talked about between publisher and followers to construct their Local Topic Space.

To distill the topics that users are interested in, documents of LDA should naturally correspond to tweets content. Since our goal is to understand the topics that each user is interested in rather than the topics that each single tweet talks about, we aggregate the tweets published by each user into one document. We adapt the original LDA model to the Twitter by replacing documents with aggregated tweet documents for every user in the local network. Hence a user can be represented as a multinomial distribution over topics corresponding to the topic distribution of the user's subjective model. The generative process can be graphically represented using plate notation in Figure 1. Formally, given a set of users U and the num-
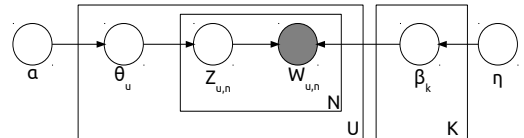

Figure 1: Plate illustration of the user-level LDA model.

ber of topics K, each user $u \in U$ could be represented by a multinomial distribution θ_u over topics with a Dirichlet prior parameterized by α. A topic k is represented by a multinomial distribution β_k with another Dirichlet prior parameterized by η. The generative process works as follows:

- For each user u, draw $\theta_u \sim Dir\left(\alpha\right)$.
- For each word $w_{u,n}$ in a user document, $n \in \{1, \cdots, N\}$:
  - draw a topic $z_{u,n} \sim Multinomial\left(\theta_u\right)$;
  - draw a word $w_{u,n}$ from multinomial probability $p\left(w_{u,n}|z_{u,n}, \beta_k\right)$ conditioned on the topic $z_{u,n}$.

The parameters θ_u and each β_k can be estimated by Gibbs sampling or variational inference. We use variational inference-based topic model package Gensim[28].

3.3.2 Sentiment Analysis of Tweets

The tweets that users have published are encoded with their opinions towards topics of interest. In order to explore the opinions of users, we need to understand sentiment embedded in each tweet, which is the task of sentiment analysis. Sentiment analysis begin with traditional text genres such as reviews and news comments. The freely expressing opinions of Twitter has made it target of sentiment analysis to understand public opinions about topics or products. But particular language characteristics of short length and informal expressions pose challenges to traditional sentiment analysis techniques. The main sentiment analysis techniques could be categorized into machine learning based approaches and rule based approaches. Machine learning based approaches often need labelled data for the training process, which is often impossible for Twitter because of the large volume and dynamic language characteristics. Rule based approaches could adapt to Twitter with good flexibility by changing the particular characteristics into rules, which is proved more robust and efficient[35, 11].

In our work, we use the rule based package SentiStrength[35]. SentiStrength has been built especially to cope with sentiment analysis in short informal text of social media. It combines lexicon-based approaches with sophisticated linguistic rules adapted to social media, which is more suitable for analyzing sentiment of tweets than other approaches in our research settings. SentiStrength assigns two values to each tweet standing for sentiment strengths: a measure of positive and a measure of negative sentiment, both on absolute integer scales ranging from 1 to 5, with 1 denoting neutral sentiment and 5 denoting highest sentiment strength.

- The positive sentiment score $p \in [1, 5]$, is basically equal to the sentiment score of the most positively classified word in the tweet, adjusted by linguistic rules.
- The negative sentiment score $n \in [-5, -1]$, is basically equal to the sentiment score of the most negatively classified word in the tweet, adjusted by linguistic rules.

Another reason we use SentiStrength lies in that sentiment value is not simple binary label but a fine-grained strength scales which is in accordance with opinion valence space of subjective model and catch fine opinion distributions of users. For the convenience of distribution calculation, we map the output of SentiStrength to single-scaled opinion valence space $[0, 8]$ based on:

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

In the opinion valence space, value 4 indicates neutral sentiment, while values above 4 indicate positive sentiment and values below 4 negative sentiment, so that we can aggregate all sentiments towards a topic as a opinion distribution over opinion valence space.

3.3.3 Concrete Subjective Model

After statistical topic analysis and sentiment analysis illustrated above, we could now concrete subjective model in a local network settings. Suppose a tweet collection published by a user u as $C_u = \{c_i|i \in [1, \cdots, N]\}$, and C_u is concatenated to one document d_u during the construction of Local Topic Space $T = \{t_i|i = 1, \cdots, K\}$. In the Local Topic Space T, a topic model is built with parameter θ representing the distribution of each user over topics his tweets talk about, and parameter β representing the distribution of each topic over the vocabulary of all tweets. To find the sentiment of each tweet c in collection C_u, sentiment analysis technique is applied to c and outputs sentiment strength s_c as the Section 3.3.2 has described. The subjective model of user u is built according to following procedure:

- Firstly, for user u, the corresponding component θ_u of parameter θ is the distribution of his topics of interest in the Local Topic Space T, and probability $p\left(z_u|\theta_u\right)$ could be regarded as the weight of subjective model $w_u\left(t_i\right)$. Topics of interest could be defined as $Z_u = \{z_u|p\left(z_u|\theta_u\right) > 0\}$.
- Secondly, as topics of each tweet c are considered as the target of sentiment expressed in c, the topic model of Local Topic Space T is applied to each tweet c and outputs topics c talks about as $Z_c = \{z_c|p\left(z_c|\theta, \beta, Z_u\right) > 0\}$.
- Thirdly, the opinion distribution of user u towards topic $t \in Z_u$ could be calculated as:

$$d_{u,t}\left(o\right) = \left\{\frac{N_o}{\sum_{o \in O} N_o}|O = [0, \cdots, 8]\right\} \quad (4)$$

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

$$N_o = \sum_{c \in C_u} I\left(s_c\right), \text{ if } s_c = o \& t \in Z_c \quad (5)$$

$$I\left(s_c\right) = \begin{cases} 1 & \text{if } s_c = o \& t \in Z_c \\ 0 & \text{else} \end{cases} \quad (6)$$

For simplicity, we assume the opinion of tweet c is expressed towards every topic it talks about in Z_c.

Totally, the subjective model of user u is thus be describe as:

$$P\left(u\right) = \{(t, p\left(z_u|\theta_u\right), d_{u,t}\left(o\right))|t \in Z_u, o \in O\} \quad (7)$$

and accordingly the subjective model of tweet c is:

$$P\left(c\right) = \{(t, p\left(z_c|\theta, \beta\right), d_{c,t}\left(o\right))|t \in Z_c, o \in O\} \quad (8)$$

3.3.4 Retweet Analysis Formulation With Subjective Model

On the psychological theory, our study try to understand the underlying reasons that a user retweet a tweet by investigating the relationship among subjective model of publisher, follower and tweet. For a tweet c, the corresponding publisher p, and a list of followers $F = \{f_i|i = 1, \cdots, N\}$, for each $f_i \in F$, a tuple $\langle f_i, p, c, r_{fpc}\rangle$ could be defined as Section 3.2. We firstly build subjective model $P\left(u\right)$ for each user $u \in F \bigcup p$ and $P\left(c\right)$ for tweet c in the Local

Table 1: Retweet Dataset Statistics

Total tweets which have been retweeted	500
Average number of followers per tweet	89
Total retweeters	5214
Total non-retweeters	40317

Topic Space T defined by users in $F \bigcup p$. We are interested in how a tweet resonates with the users who might want to retweet it. By using the word "resonate", we assume that a user retweet a message because he not only find its topics interesting but also share similar opinions towards these topics. With the subjective models built for users and tweets, we could define a similarity measurement to quantify the resonation among them:

$$Sim\left(c, f_i\right) = similar\left(P\left(c\right), P\left(f_i\right)\right) \quad (9)$$

according to Equation 7,8:

$$Sim\left(c, f_i\right) = \lambda * Dist\left(p\left(z_c|\theta,\beta\right), p\left(z_{f_i}|\theta_{f_i}\right)\right)$$
$$+ \left(1 - \lambda\right) * \left(\sum_{t \in T} Dist\left(d_{c,t}, d_{f_i,t}\right)\right) \quad (10)$$

where

- λ is coefficient used to control the proportions of topic similarity and opinion similarity in the holistic subjective similarity, and we initiate it by setting $\lambda = 0.5$.
- $Dist$ is the similarity measurement between two distribution, we use *Cosine Similarity* in our research.

We also assume that a user might retweet another user because of their subjective resonation, accordingly we could define similarity between publisher p and follower f_i as:

$$Sim\left(p, f_i\right) = \lambda * Dist\left(p\left(z_p|\theta_p\right), p\left(z_{f_i}|\theta_{f_i}\right)\right)$$
$$+ \left(1 - \lambda\right) * \left(\sum_{t \in T} Dist\left(d_{p,t}, d_{f_i,t}\right)\right) \quad (11)$$

4. EXPERIMENT

In this section, we systematically investigate how subjective model is built and its influence on retweet behavior on an off-the-shelf dataset.

4.1 Dataset

We adopt an off-the-shelf Twitter dataset[17]. For the dataset, 500 randomly selected English tweets which had been retweeted at least once were used as test tweets, then each test tweet was chosen as starting point to collect data of its publisher and followers. Summary statistics of the dataset are listed in Table 1.
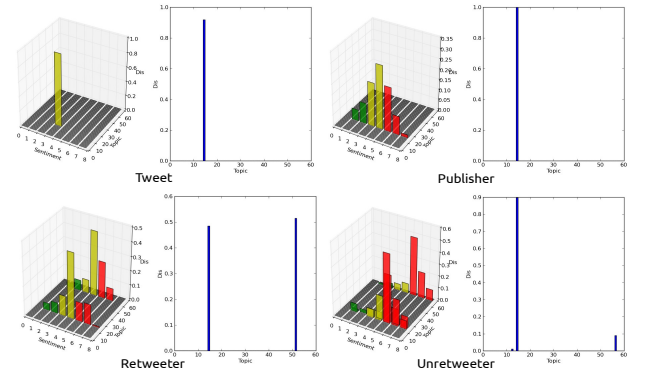
4.2 Example of Subjective Model

As the core of our work, how to build subjective model has been elaborated in section 3.3.3, in this section we give an qualitative description about subjective model and its ability in explaining the retweet behavior with an intuitive exmple. There are 500 test tweets, 500 corresponding publishers, 4,5531 followers and 6,277,736 published tweets in the dataset, the relations are illustrated in Figure 2. There is a local network structure for each tweet of 500 test tweets as figure shows, consisting of its publisher and followers connected
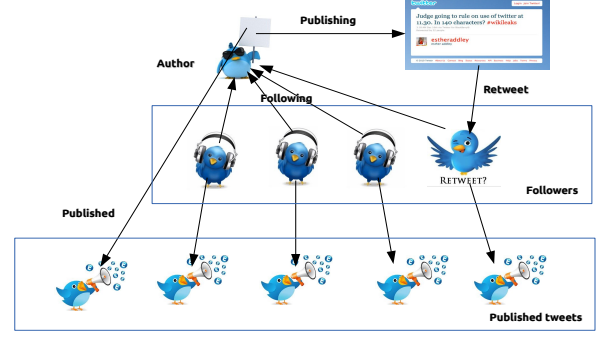
Figure 2: Illustration of dataset structure.

with "following" relationship. We build a local topic space for each local network in which subjective models of users and tweets are built. As an example, we present subjective models for one of the 500 test tweet, its publisher, and two followers(one retweet the tweet while the other does not) as Figure 3 shows, in which the right part of each model is topic distribution, the left part is opinion distribution for each topic. It is the 14th topic that the tweet talks


Figure 3: Subjective model examples.

about in the local topic space. In order to make Figure 3 vividly, we plot top words of the 14th topic , the tweets of publishers and two followers in a word cloud figure as Figure 4 shows[2]. In the example, the test tweet content is:

Tweet: *Sometimes the right person for you was there all along. You just didn't see it because the wrong one was blocking the sight.*

The tweet content is about topic of "love between people" and the opinion of tweet is neutral, which is in accordance with the 14th topic word cloud in Figure 4 and subjective model of tweet in Firgure 3. The tweets of the publisher also talks about "love between people", and his opinion is mainly neutral as it is demonstrated from Figure 3,4. As for two followers, the "retweeter" who retweet the example tweet has published tweets about two topics(the 14th and 52nd topic) uniformly and his opinions towards the two topics are mainly neutral; while the other one who does not retweet the example tweet(we call "unretweeter") has also talked about two

---

[2]We use TagCrowd(`http://tagcrowd.com/`) to produce word cloud
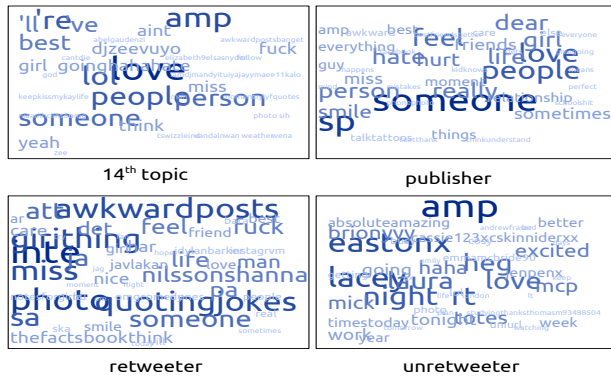
Figure 4: Word cloud of 14th topic, publisher and followers

Figure 5: Influence of different factors on retweet.

topics(14th and 56th topic), but he is mainly interested in "love between people" topic and his opinion is positive. Although two followers have common topic of interest(the 14th topic), the difference of their opinion towards the topic elicits their different retweet behavior, which proves subjective model is able to understand the underlying reasons of retweet behavior.

4.3 Influence of Subjective Model on Retweet

In this section, We quantitatively investigate the influence of subjective model on rewteet behavior with factors derived from it. In our formulation of retweet problem we model retweet with subjective model in the form of similarity measurement 10,11. By setting different value to λ, the measurement can be divided into different parts to model different factors that might influence user's rewteet behavior, which are:

- **TTF**: **T**opic similarity between **T**weet and each **F**ollower ($\lambda = 1$ in measurement 10)
- **OTF**: **O**pinion similarity between **T**weet and each **F**ollower ($\lambda = 0$ in measurement 10)
- **STF**: **S**ubjective similarity between **T**weet and each **F**ollower ($\lambda \in (0, 1)$ in measurement 10)
- **TPF**: **T**opic similarity between **P**ublisher and each **F**ollower ($\lambda = 1$ in measurement 11)
- **OPF**: **O**pinion similarity between **P**ublisher and each **F**ollower ($\lambda = 0$ in measurement 11)
- **SPF**: **S**ubjective similarity between **P**ublisher and each **F**ollower ($\lambda \in (0, 1)$ in measurement 11)

To analyze the influence of different factors on retweet, we averaged six similarity scores on 5214 followers who rewteet the test tweets and 5214 randomly selected followers who do not rewteet separately. Figure 5 shows the comparing result. As the figure demonstrated, on average, the similarities scores of retweeted followers are clearly higher than unretweeted followers for all six factors, which proves that our assumption is reasonable. Specifically,

- TTF score shows that a tweet is more likely to be retweeted by followers who find topics it talks about interesting to them, which is consistent with other studies[18, 9];
- OTF score shows that opinions in a tweet is an important indicator to be rewteeted by by followers who hold similar
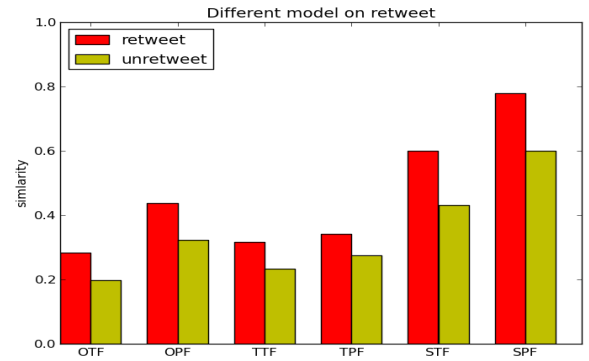
opinions, although other studies[26, 21] have shown that sentiment in tweet have impact on rewteet, most of them don't consider the opinions of followers and the opinion similarity betweet tweet and its followers;

- STF score shows the subjective model we put forward is the most distinguishable feature among the six factors with the largest difference between retweeted and unrewteeted followers, which proves the importance of subjectivity;
- TPF score gives another perspective for retweet from the topic similarity between tweet publisher and followers, indicating that followers are more likely to retweet those whose interests are similar, which verifies the homophily principle of rewteet relation;
- OPF score indicates that similar opinions for common topics of interest also influence followers' decision of retweeting anothor user, which may be another proof of homophily of rewteet relation.
- SPF score is interesting in that it implies that subjective similarity between user and follower might cause retweet, and we call this phenomenon "tight homophily" because it requires both topic homophily and opinion homophily.

The six scores used to model the factors that influence rewteet could be grouped into two aspects. One is consisted of TTF, OTF and STF, which is direct and explicit by modeling the tweet and its followers; the other is consisted of TPF, OPF and SPF, which is indirect and implicit by modeling the tweet publisher and followers. The two aspects reflect properly the information sharing and diffusion structure of Twitter at micro-level as illustrated in Figure 2.

4.4 Performance of Retweet Prediction

The main purpose of subjective model is to help users find attracting information which could arouse their resonation from the overwhelming information streams. In the context of Twitter, retweet is an important signal elicited by such resonation, because users are prone to broadcast their favorite tweets to their followers. Thus, the performance of predicting retweet is a suitable measurement for the utility of subjective model. The experiment can be regarded as a simulation of information diffusion process: when a user is browsing message streams he has subscribed for, he might find himself resonate with a tweet and retweet it to his followers. On the other side, when a new tweet is created, we want to know those followers who will retweet it after reading it.

As Section 3.2 stated, the retweet problem we study could be formulated as a tuple $< f, p, c, r_{fpc} >$. In the prediction experiment we need to determined the label r_{fpc} for each $c, p,$ and f. There are 5,214 users in our dataset who retweet a test tweet, so we could extract 5214 tuples as positive instances with their label $r_{fpc} = 1$. The other 40,317 users who do not retweet any test tweet are also extracted to form negative tuples with label $r_{fpc} = 0$. For the consideration of balance between positive and negative instances, we randomly sample 5,214 negative instances into the final dataset.

4.4.1 Comparison With Other User Models

Firstly the comparison between our model with other user models (TF-IDF model[17], entity-based model and hashtag-based model[1]) in predicting retweet are investigated. As for our model, the six parts defined above are used for comparison because they model different factors that influence retweet. For the comparing models, cosine similarities are calculated between tweets and their followers. We use the logistic regression classifier of Scikit-learn machine learning package[24] for training, with 5-fold cross-validation on our balance dataset. And accuracy measurement is used for evaluation. Figure 6 shows the results of our model and all other models. The highest accuracy of 68.67% is the STF(Subjective simi-
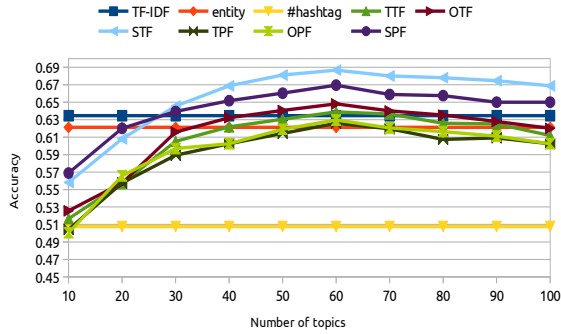

Figure 6: Comparison of different models.

larity between Tweet and Followers) model, which is clearly better than other models. The accuracies of TF-IDF model and entity-based model are 63.45% and 62.12%, which are very close to our TTF(Topic similarity between Tweet and Followers, 63.88%) model and OPF(Opinion similarity between Publisher and Followers, 62.96%) model. While for hashtag-based model, its accuracy is 50.76% , which is only little better than random selection (50%) in that a very low usage of hashtag in the dataset. The accuracies of the other three model of our method are OTF(Opinion similarity between Tweet and Followers) model 64.80%, TPF(Topic similarity between Publisher and Followers) model 62.58% and SPF(Subjective similarity between Publisher and Followers) model 66.95%. The results show that subjective model can reach a better understanding of user retweet behavior than the other models.

Figure 6 also shows that the influence of topic number of LDA on the predicting accuracy, which arrives its peak when the number is set to 60.

4.4.2 Retweet Classification Evaluation

In this section, we feed the six parts of our model as features into a retweet classification framework to verify the effectiveness of

Table 2: Prediction Accuracy of Different Models. Significant improvement over baseline with star(∗) and LUO' model with dagger(‡) (p<0.05).

Feature Set	Accuracy(%)	Significance	
RB	60.85		
LUO	68.76	∗	
SM6	69.12	∗	
LUO(⊖)+TTF	69.20	∗	
LUO(⊖)+TPF	71.04	∗	‡
LUO(⊖)+OTF	71.88	∗	‡
LUO(⊖)+OPF	70.27	∗	
LUO(⊖)+STF	72.86	∗	‡
LUO(⊖)+SPF	72.05	∗	‡
LUO(⊖)+All	72.93	∗	‡

the subjective model. We compare the performance of our model with a prediction model of Luo et al.[17] which use four feature families: Retweet History, Follower Status, Follower Active Time and Follower Interests. We use LinearSVM of Scikit-learn package to build a retweet prediction model, leveraging two different features sets, one includes the six features derived from subjective model(marked as "SM6"), and the other is Luo et al.[17](marked as "LUO") feature set in which they use "bag-of-words" to model the followers interest. We use the same dataset as Section 4.4.1 with 5-fold cross-validation, and accuracy as measurement. In addition, we set a baseline(marked as "RB") for evaluation, in which followers who have retweeted the publisher's previous tweets before are predicted as retweeters for current tweet. The result is listed in Table 2. As shown in table, the accuracy of baseline is 60.85%, and both sets of features(LUO and our SM6) outperform the baseline significantly. But our feature set shows no significant improvement over LUO feature set since the our model only tries to reflect the retweet motivation of users based on content, whereas other important factors associated with retweet are not considered, such as network context and reading habit of the user. As denoted by "LUO(⊖)+" in the table, we combine the two sets of features by replacing the Follower Interests features of LUO model with our six features one by one. The accuracies are all improved over LUO after adding our features, which means that our model is of great importance for retweet prediction models. Notice that, the most significant improvement(LUO(⊖)+STF, 72.86% versus 68.76%) is the subjective similarity features between tweet and followers, which proves our assumption that resonation between tweet and the followers elicits retweet behavior. Besides, the improvement by adding subjective similarity features between publisher and followers(LUO(⊖)+SPF, 72.05% versus 68.76%) is also obvious in that the resonation between publisher and follower indicates the tight homophily between them. And finally, the last row of table is the complete combination of two sets of features(LUO(⊖)+All) by adding all six features to LUO feature set, and the performance shows no significant improvement over adding only STF feature.

5. CONCLUSION

In this paper, we propose subjective model to analyze user retweet behavior on Twitter in the light of psychological theory by assuming that retweet behavior should be elicited by the subjective resonation between the tweet and its followers. We define subjective model formally as combination of topic distribution and opinion distribution, and we concrete subjective model leveraging statistical topic model and sentiment analysis techniques. We demonstrate the effectiveness of the proposed model with retweet analysis prob-

lem and show that our model is able to reach more comprehensive understanding of user retweet behavior.

Our future work mainly lies in two directions. Firstly, our subjective model is established in a simple way, and it is an interesting direction to establish it under the framework of generative topic-sentiment model, which has been applied in reviews and citation network. Secondly, we will apply subjective model to other social media analysis task such as connection prediction and friend recommendation.

6. REFERENCES

[1] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing user modeling on twitter for personalized news recommendations. In *Proceedings of the 19th international conference on User modeling, adaption, and personalization*, UMAP'11, pages 1–12, Berlin, Heidelberg, 2011. Springer-Verlag.

[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*, CIKM '12, pages 1602–1606, New York, NY, USA, 2012. ACM.

[3] D. Blei, A. Ng, and M. 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 *2010 43rd Hawaii International Conference on System Sciences*, volume 0, pages 1–10, Kauai, HI, 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*, KDD '11, pages 150–158, New York, NY, USA, 2011. ACM.

[6] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1185–1194, New York, NY, USA, 2010. ACM.

[7] G. Comarela, M. Crovella, V. Almeida, and F. Benevenuto. Understanding factors that affect response rates in twitter. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, HT '12, pages 123–132, New York, NY, USA, 2012. ACM.

[8] 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):1261–1268, 2007.

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

[10] J. Hannon, M. Bennett, and B. Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 199–206, New York, NY, USA, 2010. ACM.

[11] X. Hu, J. Tang, H. Gao, and H. Liu. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web*, WWW '13, pages 607–618, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[12] 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*, WSDM '13, pages 537–546, New York, NY, USA, 2013. ACM.

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

[14] M. Jenders, G. Kasneci, and F. Naumann. Analyzing and predicting viral tweets. In *Proceedings of the 22nd international conference on World Wide Web companion*, WWW '13 Companion, pages 657–664, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[15] 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*, HLT '11, pages 151–160, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[16] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[17] Z. Luo, M. Osborne, J. Tang, and T. Wang. 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, pages 869–872, New York, NY, USA, 2013. ACM.

[18] S. A. Macskassy and M. Michelson. Why do people retweet? anti-homophily wins the day! In L. A. Adamic, R. A. Baeza-Yates, and S. Counts, editors, *ICWSM*. The AAAI Press, 2011.

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

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

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

[22] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Searching microblogs: coping with sparsity and document quality. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 183–188, New York, NY, USA, 2011. ACM.

[23] M. Osborne and V. Lavrenko. RT to Win! Predicting Message Propagation in Twitter. *Artificial Intelligence*, pages 586–589, 2011.

[24] 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.

[25] M. Pennacchiotti and A.-M. Popescu. A Machine Learning Approach to Twitter User Classification. In *International AAAI Conference on Weblogs and Social Media*, 2011.

[26] R. Pfitzner, A. Garas, and F. Schweitzer. Emotional divergence influences information spreading in twitter. In

J. G. Breslin, N. B. Ellison, J. G. Shanahan, and Z. Tufekci, editors, *ICWSM*. The AAAI Press, 2012.

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

[28] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[29] 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*, CSCW '12, pages 7–16, New York, NY, USA, 2012. ACM.

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

[31] S. Stieglitz and L. Dang-Xuan. Political communication and influence through microblogging-an empirical analysis of sentiment in twitter messages and retweet behavior. In *HICSS*, pages 3500–3509. IEEE Computer Society, 2012.

[32] 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 *Proceedings of the IEEE Second International Conference on Social Computing (SocialCom)*, pages 177–184, Minneapolis, Aug. 2010. IEEE.

[33] C. Sunstein. *On Rumors: How Falsehoods Spread, Why We Believe Them, What Can Be Done*. Farrar, Straus and Giroux, 2009.

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

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

[36] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In B. D. D. 0001, T. Suel, N. Craswell, and B. L. 0001, editors, *WSDM*, pages 261–270. ACM, 2010.

[37] Z. Xu, Y. Zhang, Y. Wu, and Q. Yang. 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, pages 545–554, New York, NY, USA, 2012. ACM.

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