From Interests to Opinions: Modelling Subjectivity for Retweeting Analysis

ABSTRACT

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

Categories and Subject Descriptors

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

General Terms

Model, Experimentation

Keywords

Twitter, subjectivity, retweet, LDA, sentiment analysis

1. INTRODUCTION

It is well recognized that online social networks such as Microblogging is a complex and subtle platform that improves the diffusion of information. With the help of Microblogging, a company can market a new product by triggering and cascading a large number of users to adopt the product through the effect of "word of mouth" in the social networks. Twitter, one of the most popular Microblogging services, has become a center of attention due to the amount of users it has attracted and the volume of messages it produces. The retweeting convention and complex network of Twitter provide an unprecedented mechanism for the spread of information despite the restricted length of a single message (i.e. a tweet of 140-characters limits). From the point of micro-level, retweeting behavior allow the flow of information because it indicates situations where a user felt a tweet was important enough that he shared it with his followers. Actually almost a quarter of the tweets of a user are retweeted from other users [40]. For this reason, understanding how retweeting behavior works can help explaining information dissemination on Twitter.

For a user, retweeting is a process that includes reading the tweet, evaluating the content and deciding whether to share. The crucial part is to evaluate whether a tweet contains information interesting and agreeable to be shared. Usually a user receives a great many tweets on different topics every day, whether a tweet will be retweeted depends on the subjective choice of a user. The subjective initiative nature of human determines that his behavior pattern is subjectivity driven, and psychological researchers have identified subjectivity as the underlying factor that influences human's behaviors [24]. Also according to theory of Biased Assimilation, people tend to choose and disseminate information according to their own biased subjectivity [14]. On twitter, users are inclined to present their subjectivity by discussing various topics online and expressing their opinions toward these topics [5]. Therefore modelling the subjectivity of users will provide an important perspective for retweeting behavior analysis. This research is motivated by a desire to find what drives subjective users of social media to disseminate information they come across.

The problem can be clearly explained with Figure 1. The figure illustrates a social network consisting of users, following relations between them and their associated tweets (posted by themselves or retweeted from their followees). Users usually present their opinions by gennerating content on topics they are interested in, therefore the subjectivity of users are encoded in the tweets they have generated. In our example, the tweets of the users present their different opinions about two topics: cellphone "Iphone" and movie "Frozen" (with the color "red" standing for positive evaluation, "green" for negative, "yellow" for neutral.). Tony and Jane were positive about movie "Frozen", while Ada was negative and Yang was neutral. The problem we study here can be described as: now Tony posts a new tweet which is positive about "Frozen", we want to find who is more likely to rewteet it among his three follwers considering their subjective preferences?

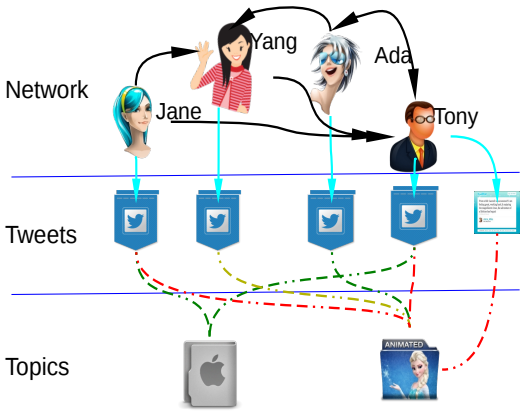

Figure 1: Motivating example.

Intuitively, based on the principle of "like attracts like", a biased user is more prone to retweet a message that meets his own subjectivity tastes. Therefore there are two questions arising to solve the problem: how to model the subjectivity information for users and tweets, and how to measure the similarity for the subjectivity information? Answering the questions is non-trivial. Indeed, it is challenging on the following aspects.

Firstly, how to efficiently mining latent topics and topic-related opinions to model the subjectivity? There are hundreds of millions of users discussing various topics on Twitter. How to identify topics a target user is interested in and mine his opinions toward these topics efficiently from User-Generated Content (UGC) is a main challenge. In particular, when Twitter is a heterogeneous social network (consisting of heterogeneous objects such as users and tweets) [17]. Thus besides the network structure, the content spreading on the top of networks becomes a key factor for topic and opinion mining in heterogeneous networks.

Secondly, how to define effective opinion representation and subjectivity similarity to evaluate the agreement of two subjective objects? Most previous opinion mining researches [19] use three scalars (+1, 0, -1) to describe users' opinion or sentiment towards different topics: +1 means opinion agreement or positive sentiment, -1 means opinion disagreement or negative sentiment, and 0 means no opinions or neutral sentiments. It can not distinguish the difference between two subtle opinions, for example, when one is strongly positive and the other is weakly positive. Therefore a more fine-grain representation is needed to describe users' opinions. Besides, opinion toward a topic is not always the same when considering different aspects of the topic, so it is better to describe users' topic preference as a probability distribution over the sentiment valence space. Then how to define and calculate the subjectivity similarity with such a special opinion representation? This problem has not been considered before.

Thirdly, how to demenstrate the correlation between the subjectivity and the users' retweeting behavior? Is it true that more similar between a tweet and a user, he will be more likely to retweet it? If so, to what extent can subjectivity improve the retweeting analysis performance? A systematic investigation of these problems is still needed.

There have been many studies trying to identify factors that influence whether a tweet will be retweeted [4, 17]. However few studies have investigated the subjective motivation of a user to retweet a message. Previous studies on retweeting analysis have shown that an enriched user model gives coherent and consistent explanation for retweeting analysis [21, 9]. Specifically, researchers have tried to model users from four types of information: profile features ("**Who you are**"), tweeting behavior ("**How you tweet**"), linguistic content ("**What you tweet**") and social network ("**Whom you connect**") [29]. Especially, interests of a user, i.e. topics encapsulated in UGC, have been proved consistently dependable for behavior analysis [30]. However, to our best knowledge, few studies have considered the subjective aspect ("**what's your opinions**") when modelling a user. In this paper, we propose a novel method to model subjectivity of users and tweets as well (defined as subjectivity model) by combining both the topics and opinions.

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

- In the light of psychological theory, we firstly put forward formal definition of subjectivity model which incorporate topic modelling, and sentiment analysis into one unified model.
- Based on a fine-grain opinion representation, we put forward a novel way to measure the subjectivity similarity, which can distinguish subtle opinion difference between two subjective objects.
- We systematically evaluate the impact of subjective model on retweeting behavior. Expertiment results show that retweeting behaviors are correlated with subjectivity similarity, the subjectivity model outperforms topic-based model for retweeting prediction, and the performance of an off-the-shelf predicting model is significantly improve by combining with our model.

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

2. RELATED WORK

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

2.1 Retweeting Analysis

A large body of studies have analyzed characteristics of retweeting, examining factors that lead to increased retweetability and designing models to estimate the probability of being retweeted. Suh *et al.* [37] found that tweets with URLs and hashtags were more likely to be retweeted. Macskassy and Michelson [21] found that models derived from tweet content could explain most of retweeting behaviors. Comarela *et al.* [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. Starbird and Palen [35] addressed specifically the retweeting mechanism during crises and found that tweets with topical keywords were more likely to be retweeted. Osborne and Lavrenko [30] 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 to predict retweeting. Jenders *et al.* [15] analyzed the "obvious" and "latent" features from structural, content-based, and sentimental aspects of both tweets and users, with respect to their impact on the spread of tweets. Naveed *et al.* [26, 25] introduced interestingness of tweets, and quantified it based on such features as emoticons, sentiments and topics to predict the probability of retweet for an individual tweet. Feng and Wang [9] built a graph with all sources of information incorporating into nodes and edges, and proposed a feature-aware factorization model to rerank the tweets according to their probability of being retweeted. Pfitzner *et al.* [31] proposed a new measure called emotional divergence to evaluate the retweet probability of a tweet and showed that highly emotional diverse tweets can have up to almost five times higher chances of being retweeted.

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

2.2 Sentiment Analysis

Sentiment analysis is a popular research area for years. Previous research mainly focused on reviews or news comments [27]. Recently, researchers began to pay more and more attention to social media such as Microblogging. Hu *et al.* [12] interpreted emotional signals available in social media data for unsupervised sentiment analysis by providing a unified way to model two main categories of emotional signals: emotion indication and emotion correlation. Jiang *et al.* [16] focused on target-dependent Twitter sentiment classification, they proposed a method to improve target-dependent Twitter sentiment classification by taking target-dependent features and related tweets into consideration. Asiaee T. *et al.* [2] presented a cascaded classifier framework for per-tweet sentiment analysis by extracting tweets about a desired target subject, separating tweets with sentiment, and setting apart positive from negative tweets. Hu emphet al. [13] extracted sentiment relations between tweets based on social theories, and proposed a novel sociological approach to utilize sentiment relations between messages to facilitate sentiment classification and effectively handle noisy Twitter data. Motivated by sociological theories that humans tend to have consistently biased opinions, Calais Guerra *et al.* [5] addressed challenges of topic-based real-time sentiment analysis by proposing a novel transfer learning approach with a suitable source task of opinion holder bias prediction. Thelwall *et al.* [39, 38] designed SentiStrength, an algorithm for extracting sentiment strength from informal English text by exploiting the grammar and spelling styles in typical social media text. In this paper, we adopt SentiStrength for sentiment analysis to build our subjective model, as a more fine-grain sentiment strength could give us more detailed opinion of users than binary polarized sentiment.

2.3 Topic-Sentiment Model

Since the introduction of LDA model [3], various extended LDA models have been developed for topic extraction from large-scale corpora at user level [34, 32]. Topic models can also be utilized in sentiment analysis to correlate sentiment with topics. Topic Sentiment Mixture (TSM) model [23] represents the sentiment as a language model separated from topics, which means TSM considers the topic and sentiment orthogonally, the word samples from either topics or sentiments. Multi-Aspect Sentiment (MAS) model [41] aims at modeling topics to the predefined aspects that are explicitly rated by users in reviews, from which the sentiment is modeled on the aspect level according to the sentiment distribution from a weighted combination from extracted topics and words. Joint Sentiment/Topic (JST) model [18] presents a novel way to detect the sentiment of document with topic extraction and its sampling process considers that the topics are associated with sentiment and document, which can model the topic and sentiment simultaneously. These models are similar with our work in mining topic-related opinion at document or user level. But all of them try to use a general word-sentiment distribution to model the sentiment of blogs or reviews. However, social media users often show their sentiment with special spellings beyond such general word list, which forms the specific characteristics of social media sentiment expressions. In fact, rule-based sentiment analysis methods can catch some subtle sentiment of tweets by transforming these characteristics into rules. Therefore, we construct our model in a framework analyzing topics and opinions separately but not in a unified generative way as TSM and JST.

3. SUBJECTIVITY MODEL

Subjectivity has been extensively studied by psychologists to characterize the personality of a person based on his historical behaviors and remarks [8]. Linguists define the subjectivity of language as speakers always show their perspectives, attitudes and sentiments to events, people, topics, and entities in their linguistic contents [36]. However, how to computationally model the subjectivity of a user is still an open challenge. The advent of online social media such as Twitter has given a new layout to the challenge. Twitter allows users to show their personal subjectivity by publishing short messages, which provides researchers with data resources to model the subjectivity of users. Therefore, we give a formal definition of the subjectivity model under the context of Twitter.

3.1 Intuitions and Our Approach

On twitter, users are often interested in different topics, e.g., Tony and Jane are interested in topics "Iphone" and "Frozen". The degrees of interests may also vary with different topics. Furthermore, the opinion of a user may not always be the same when he has posted multiple tweets on different aspects of the same topic. Thus, to summarize, we have the following intuitions for the subjectivity model:

- Each user u is associated with a vector $W_u \in R^T$ of T-dimensional topic distribution ($\sum_t w_u(t) = 1$). Each element $w_u(t)$ is the interest weight of the user on topic t.
- Opinion of each user on one topic can be represented as a vector $d_{u,t} \in R^S$ of S-dimensional sentiment distribution ($\sum_s d_{u,t}(s) = 1$). Each element $d_{u,t}(s)$ is the opinion probability of the user with setiment strength s.

Therefore, from the technique aspect, our objective is to design a method to learn the user interests (the associated topic distribution) and to estimate the opinions (the associated sentiment distribution) on different topics simultaneously. In this paper, we propose a topic-sentiment separated modelling framework. First, by combining both textual information and link information in social networks, we use a probabilistic generative model (LDA) to learn user interests which are represented as topic distributions. Second, based on the tweet-level topic and sentiment analysis, we propose an opinion integration process to derive opinions of users.

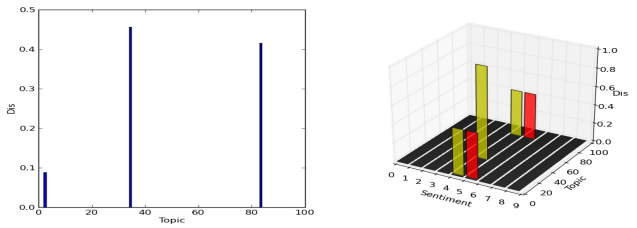

Figure 2: Subjectivity model example. The left subgraph denotes interests distribution on topic 2, 32 and 83: $(w_u(2) = 0.08, w_u(32) = 0.48, w_u(83) = 0.44)$**. The right subgraph denotes opinions towards topics:** $O_2 = (d_{u,2}(4) = 0.5, d_{u,2}(5) = 0.5)$**,** $O_{32} = (d_{u,32}(4) = 1.0)$**,**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

3.2 Definition

Let $G = (V, E)$ denote a social network on Twitter, where V is a set of users, and $E \subset V \times V$ is a set of follow relationships between users. For each user $u \in V$, there is a tweets collection M_u denoting his message history. We assume that there is a topic space T containing all topics users in V talk about, and a sentiment valence space S to evaluate their opinions towards these topics. For the "subjectivity" of a user $u \in V$, we refer to both topics and opinions articulated in his tweets collection M_u.

DEFINITION 1 (SUBJECTIVITY MODEL). *The subjectivity model $P(u)$ of user u, is the combination of topics $\{t\}$ the user talks about in topic space T and his opinions $\{O_t\}$ towards each topic distributed over sentiment valence space S.*

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

where:

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

Figure 2 shows a visualized subjectivity model of a user in a $[0, 100]$ topic space and a $[0, 8]$ sentiment valence space.

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

Our definition of subjectivity model is quite similar with existing works of topic-sentiment model [23, 18, 41]. Usually they learn a general word-sentiment distribution to model the sentiment of blogs or reviews, which may not work well for short and informal social media languages such as tweets. Compared to the traditional topic-based representation, sentiment representation is deemed to be more challenging as sentiment is often embodied in subtle linguistic mechanisms such as the use of sarcasm or incorporated with highly domain-specific information. Intuitively, sentiment is dependent on contextual information, such as language usage characteristics. Sentiment of tweets is determined not only by formal words but also by various special characteristics of Twitter languages such as emoticon, capitalized words, repeated letters and exclamation mark, etc. Those features can not be easily modeled by a probabilistic distribution. Moreover, the sentiment is often evaluated with positive, negative or neutral polarities, which can not distinguish subtle opinion difference. So a more fine-grain measurement is needed. In our model we use a discrete sentiment valence space, which can catch more subtle opinions of users. We also use a rule-based sentiment analysis method to make up the deficiency of sentiment representation of topic-sentiment model during the establishment of subjectivity model.

3.3 Establishment of Subjectivity Model

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

3.3.1 Topic Analysis

Topic analysis for all users in a global network on Twitter is a non-trivial task. There are hundreds of millions of users and billions of tweets associated with these users. The effectiveness and efficiency of the topic analysis algorithm is a challenge. However, the follow relationship on Twitter is a strong indicator of a phenomenon called "homophily", which has been observed in many social networks [22]. Homophily implies that a user follows another user because of sharing common interests. According to the principle of homophily, we put forwards the concept of **local topic space** by combining topic analysis with network topology on Twitter:

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

Previous studies have tried to identify topics from tweets by finding key words [6], extracting entities [1] or linking tweets to external knowledge categories [21]. However, works show that topic model such as Latent Dirichlet Allocation (LDA) [3] is more effective in identifying topics from short and informal social media language [11]. Therefore we adopt the user-level LDA model for topic analysis, which regards all tweets of a user as one document of LDA. The LDA model is adapted to our local topic space assumption, and the relatively tiny size and topic concentration of users in an ego network lower the impact of data sparsity, and degrade the computational difficulty of LDA.

3.3.2 Opinion Mining

In the Natural Language Processing domain, opinion mining or sentiment analysis is formally defined as the computational study of sentiments and opinions about topics expressed in a text [19]. Opinions are often regulated as sequential discrete values to represent sentiment strength. Researches on the sentiment analysis of social media have provided effective techniques and tools [39, 12]. In this work,we just make use of the off-the-shelf work, i.e. SentiStrength [39]. SentiStrength assigns two values to each tweet standing for sentiment strengths: a negative value within $[-5, -1]$ denoting negative strength, and a positive value within $[1, 5]$ denoting positive strength. The $[-5, 5]$ sentiment valence space can be

used to catch fine opinion distributions in the subjectivity model. For the convenience of calculation, we map the output of SentiStrength to a single value in sentiment valence space $[0, 8]$ as follows:

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

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

3.3.3 Concreting Subjectivity Model

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

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

$$O_t = \left\{ d_{u,t}(s) = \frac{N_s}{\sum_{s \in S} N_s} | s \in S \right\} \quad (3)$$

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

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

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

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

The whole establishment process can be summarized as Algorithm 1.

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

Algorithm 1 Establishment of subjectivity model .

Require:
The user set of a local network, U;
The tweet set published by each user u, $\{M_u\}$;
Ensure:
The subjectivity model for each user u, $P(u)$;
1: Topic analysis with a user-level LDA as Section 3.3.1, getting a topic model $P(\theta, \beta|M_u, U)$;
2: **for all** tweet $m \in M_u$ **do**
3: Sentiment analysis as Section 3.3.2, outputting sentiment of m, s_m;
4: **end for**
5: **for** user $u \in U$ **do**
6: the topic distribution is the corresponding component of parameter θ, θ_u;
7: the topics u tweets about are $Z_u = \{t|p(t|\theta_u(t)) > 0, t \in T\}$;
8: **end for**
9: **for** tweet $m \in M_u$ **do**
10: topics of m can be identified by the topic model, $Z_m = \{t|p(t|\beta, Z_u) > 0, t \in T\}$;
11: **end for**
12: **for** each topic $t \in Z_u$ **do**
13: **for** sentiment value $s \in S$ **do**
14: count the number of tweets which talk about topic t with sentiment value s, $N_s = \sum_{m \in M_u} I(s_m), if s_m = s \& t \in Z_m$;
15: **end for**
16: calculating opinion towards topic t, $O_t = \left\{ \frac{N_s}{\sum_{s \in S} N_s} \right\}$;
17: **end for**
18: establishing subjectivity model of user u, $P(u) = \left\{ \left(t, p(t|\theta_u(t)), \left\{ \frac{N_s}{\sum_{s \in S} N_s} \right\} \right) | t \in Z_u, s \in S \right\}$;
19: **return** $P(u)$;

4. RETWEETING ANALYSIS WITH SUBJECTIVITY MODEL

Apart from the context constraints such as network topology, a tweet is more likely to be retweeted by a user who finds its content worth to. Therefore, we are not interested in modelling the tweet by itself as other researchers [26, 31], but understanding the underlying reasons that a user disseminates the tweet based on his subjective initiative. We assume that if a tweet is published by the author, all followers will read it in time. Under such assumption, we investigate the problem within a 1-ego network for the author of target tweet. In the ego network, the relations among a tweet, the author and followers are illustrated as Figure 3.

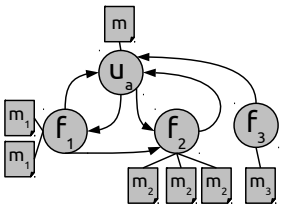

Figure 3: Illustration of relations among tweet, author and followers. Author is denoted as u_a, tweet as m, followers as f_i and tweets of follower f_i as m_i. An directed edge (f_i, u_a) means that f_i is exposed to the messages published by u_a.

4.1 Problem Formulation

The retweeting analysis problem can be formulated as following: For a target tweet m, let F denote the followers who receive m by following its author u_a, and for each user $u \in F \cup \{u_a\}$, let

M_u denote a tweet collection u has published. For each follower $u \in F$, we can define a quadruple $< u, u_a, m, r_u >$:

- r_u is a binary label indicating if m is retweeted by u.
- Firstly our work focuses on building subjectivity model $P(u)$ for each user $u \in F \cup \{u_a\}$ in the ego network with all tweets collections $M = \{M_u | u \in F \cup \{u_a\}\}$.
- Then we investigate the relation between the subjectivity of a user and his retweeting behavior to predict r_{fu} by calculating subjectivity similarities between tweet m, its author u_a and follower u.

4.2 Subjectivity Similarity

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

4.2.1 Opinion Similarity

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

$$sim^t_{opinion}(O_t^1, O_t^2) = \frac{8 - |\sum_{i=0}^{8} d_i^1 v_i - \sum_{i=0}^{8} d_i^2 v_i|}{8} \quad (6)$$

where d_i denotes the i^{th} entry of opinion distribution vector, and v_i denotes corresponding sentiment strength value. The similarities of

Table 1: Illustration of opinion similarity

	0	1	2	3	4	5	6	7	8
O_t^1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
O_t^2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
O_t^3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

opinions in Table 1 calculated with Equation 6 are $sim(O_t^1, O_t^3) = 0$, $sim(O_t^2, O_t^3) = 7/8$ and $sim(O_t^1, O_t^2) = 1/8$, which are consistent with intuitive understanding.

4.2.2 Subjectivity Similarity

As the subjectivity model indicates, a user may be interested in several topics and the weights of interests is a distribution over all topics.Therefore, the subjectivity similarity between two subectivity models can be calculated by integrating the weights of a user's topic interest and the opinion similarities on each topic. Accordingly, overall sbjectivity similarity between two subjectivity models can be calculated as Equation 7.

$$sim_{sub}(u_1, u_2) = \frac{\sum_{t=1}^{|T|} \theta_{u_1}(t) sim^t_{opinion}(O_t^1, O_t^2)}{\sum_{t=1}^{|T|} \theta_{u_1}(t)} \quad (7)$$

where T denotes the common topics between two subjectivity models, which can be regarded as the intersection between their topic sets Z_{u_1} and Z_{u_2} described in the section of subjectivity model establishment; $\theta_{u_1}(t)$ denotes the topic t weight of user u_1, and $\sum_{t=1}^{|T|} \theta_{u_1}(t)$is the normalized factor.

Note that, the subjectivity similarity is asymmetric because when we measure how similar user u_2 is with user u_1 we use the weights of common interests of user u_1. The intuition lies in that subjectivity of a user is a personal inner interest and taste, a like-minded person must resonate with the user within his interests. For our measurement of subjectivity similarity, $sim_{sub}(u_1, u_2) \neq sim_{sub}(u_2, u_1)$.

Accordingly, when we want to evaluate how similar a tweet is with a user in terms of subjectivity, we can quantitativly measure the subjectivity similarity between the tweet and the user:

$$sim_{sub}(u_1, m) = \frac{\sum_{t=1}^{|T|} \theta_{u_1}(t) sim^t_{opinion}(O_t^1, O_t^m)}{\sum_{t=1}^{|T|} \theta_{u_1}(t)} \quad (8)$$

4.3 Retweeting Analysis

The motivation of retweeting behavior is complicated, which involves the target tweet, its author and followers who is following its author, with their relations illustrated as Figure 3. The idea behind this work is that taking opinions towards interests into account can yield benefits in explaining the subjective motivation of retweeting behavior. Specifically, given a tweet m, the author u_a and any one of the followers u, we consider the probability of user u to rewteet m from three aspects:

- how similar is the tweet m to the subjectivity of user u in terms of topics and opinions, i.e.

$$sim_{sub}(u, m) = \frac{\sum_{t=1}^{|T|} \theta_u(t) sim^t_{opinion}(O_t^u, O_t^m)}{\sum_{t=1}^{|T|} \theta_u(t)} \quad (9)$$

- how like-minded are the author u_a and user u considering their similarity of subjectivity, i.e.

$$sim_{sub}(u, u_a) = \frac{\sum_{t=1}^{|T|} \theta_u(t) sim^t_{opinion}(O_t^u, O_t^{u_a})}{\sum_{t=1}^{|T|} \theta_u(t)} \quad (10)$$

- how original is the tweet m judged from its similarity with the subjectivity of its author u_a, i.e.

$$sim_{sub}(u_a, m) = \frac{\sum_{t=1}^{|T|} \theta_{u_a}(t) sim^t_{opinion}(O_t^{u_a}, O_t^m)}{\sum_{t=1}^{|T|} \theta_{u_a}(t)} \quad (11)$$

From the point of motivation, a user might retweet a message if its content is approximate to his subjectivity, its author is a like-minded friend and it is original from inner subjectivity of its author.

In next section we carry out a set of experiments to inspect and verify the impact of such motivation on retweeting behavior.

5. EXPERIMENTS

5.1 Dataset and Settings

We adopt the Twitter dataset of previous work [20]. To form the dataset, 500 target English tweets published from September 14th, 2012 to October 1st, 2012 were monitored to find who would retweet it in the next days. Besides, each target tweet was set as starting point to collect at least 200 historical tweets for its author and followers. Summary statistics of the dataset are listed in Table 2. Overall, there are 45,531 users who have posted at least 6,277,736 historical tweets. All users in the evaluation dataset were separated into the 1-ego network of their target tweet's author to establish their subjectivity model with their historical tweets. 5214 of all users retweet at least one target tweet during the monitored period. To avoid the bias introduced by dataset imbalance, an evaluation dataset was constructed by taking 5,214 retweeters as positive instances, and randomly sampling 5,214 non-retweeters as negative instances.

Table 2: Retweet Dataset Statistics

Total tweets which have been monitored	500
Average number of followers per tweet	89
All followers	45,531
All historical tweets	6,277,736
Total retweeters	5,214
Total non-retweeters	40,317

For the topic model of LDA, we use variational inference-based topic model package Gensim [33], which adopts an efficient batch-based online inference algorithm and can easily adapt to new document. All parameters are set as defaults and the number of topic traverses from 50 to 200.

5.2 Correlation Test

First of all we want to assess the existence of a correlation between subjectivity similarity and retweeting behavior. To verify such correlation, a statistical hypothesis test called Analysis of Variance (ANOVA) [10] is used. ANOVA tests the *null hypothesis* that samples in two or more groups are derived from the same population by estimating the variance of their means. This test fits our goal of testing whether the retweeters and non-retweeters have the same subjectivity similarity means. ANOVA test produces two output values: the *F-ratio* and the *p-value*. If the difference between the means is due to chance, the expected value of the *F-ratio* is 1.00, otherwise it is larger than 1.00. If the p-value is lower than the significance level α, the *null hypothesis* is rejected, which means the results is considered statistically significant. The significance level is conventionally used at 0.01. At the same time, we carry out the test by varying the topic number of LDA for topic analysis as 50, 100, 150 and 200 to determine the impact of topic number. The results are listed in Table 3. The bold-faced entries mean that the *p-value* is lower than significance level $\alpha = 0.01$.

Note that for the topic numbers of 100 and 150, all similarities yield *p-values* below α with *F-ratio* above 1.00. This suggests that the subjectivity similarities could be useful features for modeling retweeting behavior. For the rest experiments, we set the topic number as 100 for LDA model.

Table 3: ANOVA results for subjectivity similarities

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

5.3 Case Study

In this section, we give an vivid example to illustrate the subjectivity model and its ability in explaining the retweet behavior. The subjectivity models established from one of the 500 target tweets, its author, and two followers (one retweeter, the other non-retweeter) are shown as Figure 4. The right part of each sub-figure illustrates topic distribution and the left part illustrates opinions towards each topic. It is the 14^{th} topic that the tweet talks about in the local topic space. Figure 5 shows top words of the 14^{th} topic, the tweets of the author and two followers in word cloud diagrams[1].

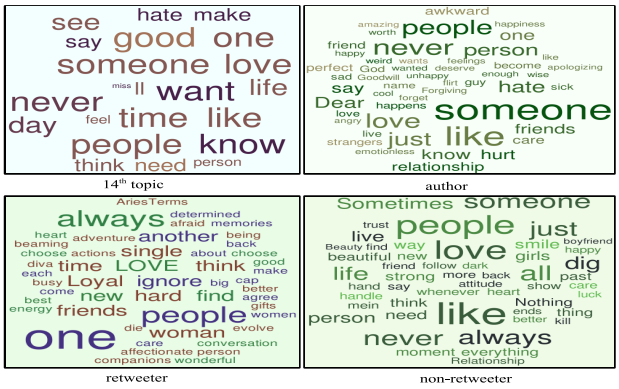

Figure 5: Word cloud diagrams of the 14^{th} topic, author and followers.

Content of the tweet 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 topic of this tweet is about "love between people" and the opinion is neutral, which is in accordance with the 14^{th} topic word cloud in Figure 5 and subjectivity model of tweet in Firgure 4. The author concentrates on the 14^{th} topic with 208 tweets, and his opinion is mainly neutral as Figure 4, 5 demonstrate ($O^{14}_{u_a} = (0, 0.04, 0.05, 0.25, 0.35, 0.25, 0.05, 0.01)$). As for two followers, the retweeter has published 250 tweets about two topics (the 14^{th} and 52^{nd} topic) uniformly (with $w_{u_r}(14) = 0.48$) and his opinion towards the 14^{th} topic is ($O^{14}_{u_r} = (0, 0.02, 0.04, 0.15, 0.50, 0.13, 0.15, 0.01)$). While the other one, the non-retweeter has also talked about two topics (14^{th} and 56^{th} topic) with 188 tweets, but he is mainly interested in the 14^{th} topic (with $w_{u_n}(14) = 0.98$) and his opinion is positive ($O^{14}_{u_n} = (0, 0.01, 0.04, 0.10, 0.25, 0.45, 0.13, 0.02)$).

Table 4 shows the three subjectivity similarities for both retweeter

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

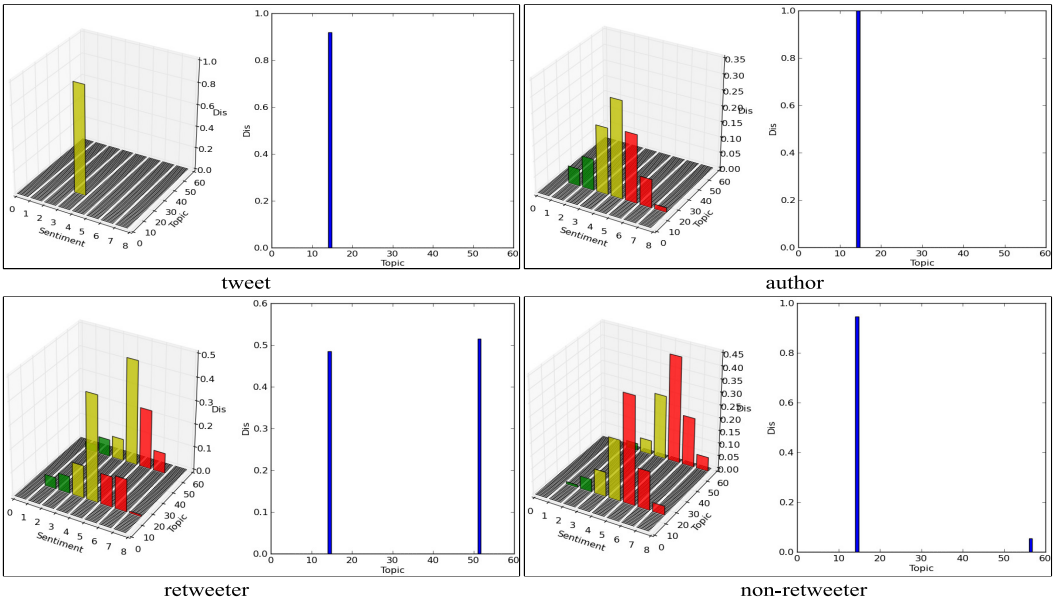

Figure 4: An illustration of subjectivity models of a tweet, author and two followers.

and non-retweeter. It is clear that except for the similarity between the tweet and its author, the other two subjectivity similarity scores of the retweeter are larger than the non-retweeter. Two followers have same interest (the 14^{th} topic), and the non-retweeter is more similar with the tweet and its author than the retweeter in interests. But their different opinions elicit their different decision, which verifies subjectivity model can help better understanding the retweeting behavior not only from topics but also opinions.

Table 4: Illustration of example subjectivity similarities

Similarity	$sim_{sub}(u, m)$	$sim_{sub}(u, u_a)$	$sim_{sub}(u_a, m)$
Retweeter	0.854	0.967	0.886
Non-retweeter	0.805	0.919	0.886

5.4 Performance Evaluation

To evaluate the performance of retweeting behavior prediction, we firstly compare our model against other topic-based models including TF-IDF model (modelling user interests with bag-of-words), entity-based model (modelling user interests with entities extracted from the UGC) and hashtag-based model (modelling user interests with hashtags used in the UGC) [1]. The cosine distance is used as similarity measurement for these models.

Secondly, our model is compared with two topic-sentiment models (TSM model [23] and JST model [18]). TSM and JST can also model topic and topic related sentiment simultaneously. But our model can capture more subtle and fine-grain opinions, which could distinguish different subective motivation of retweeting behavior. We also use Equation 7 to calculate the sentiment similarity between two users for TSM and JST. We set topic number of them to 100 in the experiments. The symmetry Dirichlet prior α and β were set to $50/T$ and 0.01 respectively for all models. The asymmetry sentiment prior γ empirically was set to (0.01, 1.8) for JST. Results of JST were averaged over 5 runs with 2000 Gibbs sampling iterations.

Finally, subjectivity model tries to catch the subjective motivation of users based on their UGC, whereas other important factors associated with retweeting behavior are not considered, such as network topology and metadata of users. Therefore, our model is also compared with the method of Luo *et al.* [20] (marked as "LUO"), in which diffenent factors that might affect rewteeting behaviors have been considered. In their work they use four feature families: "Retweet History"(follower who have retweeted a user before is likely to retweet again), "Follower Status"(the number of tweets, followers, friends, listed times and verified state), "Follower Active Time"(interaction with other users) and "Follower Interests"(TF-IDF bag-of-words model for user interests). Based on the results of comparative experiment, we also carry out combining experiments to demonstrate that performance of their method can be improved by using our model instead of bag-of-words model.

Table 5: Accuracy performance. A significant improvement over baseline with $*$ and LUO's model with ‡ ($p < 0.05$).

Feature	Accuracy(%)	Feature	Accuracy(%)
RB	60.85		
TF-IDF	62.85 $*$	LUO	71.76 $*$
entity	68.76 $*$	LUO+entity	72.15 $*$
hashtag	59.12	LUO+hashtag	68.44 $*$
TSM	67.44 $*$	LUO+TSM	68.23 $*$
JST	68.13 $*$	LUO+JST	70.53 $*$
$sim_{sub}(m, u)$	73.88 $*$ ‡	LUO+$sim_{sub}(m, u)$	74.04 $*$ ‡
$sim_{sub}(u_a, u)$	70.04 $*$	LUO+$sim_{sub}(u_a, u)$	70.27 $*$
$sim_{sub}(m, u_a)$	69.64 $*$	LUO+$sim_{sub}(m, u_a)$	71.86 $*$
sim_{all}	**75.64** $*$ ‡	LUO+sim_{all}	**78.15** $*$ ‡

The evaluation dataset is randomly divided into five parts for 5-fold cross-validation. The logistic regression classifier of Scikit-learn machine learning package [28] is used for training and testing. It is noted that followers who previously had a history of retweeting might do this in the future, so we set a baseline (marked as "RB"), which simply predicts users who have retweeted the author previously as the retweeters of target tweet. The accuracy is taken as our evaluation metric, and the results are listed in Table 5, in which

the compariative results are listed in the left part and the combining results in the right part.

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

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

Thirdly, the performance of two topic-sentiment models (TSM: 67.44%, JST: 68.13%) is not as good as our models. The reason lies in that they use a coarse-grain sentiment representation (positive or negative), which can not differentiate opinions in the same sentiment porlarity.

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

The results above show that subjectivity model can better help predicting retweeting behavior than other models and can be regarded as a better way to model the users for retweeting behavior analysis.

6. CONCLUSION

Motivated by the psychological research, this paper postulates that the online behaviors of social media users are affected by their subjectivity. Therefore, a novel subjectivity model has been proposed by combining topics and opinions to model the subjectivity of the users and tweets as well. Also an algorithm has been designed to establish the subjectivity model. To make the algorithm more efficiently, only the users of an ego network are considered and a local topic space is proposed according to the homophily principle. A novel subjectivity similarity measurement is put forward in terms of topic similarity and opinion similarity. The subjectivity model has been applied to the retweeting analysis with three subjecivity similarities among tweets, authors and followers. Experiment results demonstrate the effectiveness of the proposed model in the retweeting analysis problem and show that subjectivity model is able to reach better understanding of retweeting behavior.

In the future, we will apply the subjectivity model to other social network analysis task such as link prediction and friend recommendation.

7. REFERENCES

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

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

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

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

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

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

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

[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, 2007.

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

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

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

[12] X. Hu, J. Tang, H. Gao, and H. Liu. Unsupervised sentiment analysis with emotional signals. In *Proc. of the 22nd WWW*, pages 607–618. International World Wide Web Conferences Steering Committee, 2013.

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

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

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

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

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

[18] C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384. ACM, 2009.

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

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

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

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

[23] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180. ACM, 2007.

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

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

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

[27] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[28] 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.

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

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

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

[32] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics, 2009.

[33] 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.

[34] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

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

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

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

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

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

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

[41] T. Zhao, C. Li, Q. Ding, and L. Li. User-sentiment topic model: refining user's topics with sentiment information. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, page 10. ACM, 2012.