Cognitive Computation manuscript No.
(will be inserted by the editor)

Toward Good Information Sources: Modeling Subjectivity of Social Media User for Followee Recommendation

Songxian Xie · **Jintao Tang** · **Ting Wang**

Received: date / Accepted: date

Abstract Retweeting is the core mechanism of information diffusion on Twitter, and many factors have been proved to influence retweeting behavior, however few studies have investigated the subjective motivation of a user to retweet a message. Subjective nature of human is the underlying reason of diverse social behaviors including information diffusion, and subjective resonance triggered by topics and opinions similarity between tweets and users will elicit retweeting. In this paper, in the light of psychological theory, we assume that a tweet is more likely to be retweeted by a user because of similar subjectivity, and propose a subjectivity model to combine both the topics and opinions to model subjectivity. With state-of-the-art topic model and sentiment analysis techniques, we establish subjectivity model by finding topics and determining opinions towards these topics from user-generated content simultaneously. We evaluate our model in the retweeting analysis problem to verify its impact on retweeting and effectiveness in the retweeting prediction performance. Specifically, we demonstrate that subjective similarity is the most distinguishable feature with largest difference between retweeted and unretweeted users; subjectivity model outperforms other models in rewteeting prediction; and features derived from subjectivity model give the most significant improvement over a off-the-shelf predicting model in a classification framework.

Keywords Twitter · subjectivity · retweeting behavior · LDA · sentiment analysis

Songxian Xie
school of computer, National University of Defense Technology
Tel.: +86-0731-84574627
E-mail: xsongx@nudt.edu.cn
Jintao Tang
Jttang@nudt.edu.cn
Ting Wang
tingwang@nudt.edu.cn

1 Introduction

They provide access to a very vast source of information on an unprecedented scale. *De facto*, an online social network (OSN) can be described as a user-generated content system that permits its users to communicate and share information. Messages are the main information vehicle in such services. Users publish messages to share or for-ward various kinds of information, such as product recommendations, political opinions, ideas, etc. More data about social behavior is now available than ever before. These data, much of which come from social media sites such as Twitter, contain traces of individual activity and social interactions. On Twitter, interactions include users posting short text messages, called tweets, and following other users to receive their posts. Users may also respond to posts shared by others, for example, by retweeting them to their own followers. These abundant data offer new opportunities for learning models of user behavior and interests, identifying communities of like-minded users, and inferring the topics of conversations between them. The models can, in turn, be used to understand how popular opinion is changing, predict future activity, and identify timely and interesting information. Researchers have developed a variety of probabilistic meth- ods to learn user models from social data [13,28,76,55]. Such models usually include a users interest in some topic as a hidden parameter, which is estimated from her response to messages on that topic. The more a user responds, for example, by retweeting a message on the topic, the more she is interested in it. These models, however, fail to account for details of user behavior that affect response, such as how often the user visits Twitter, how many messages she receives, and how many of these she inspects. Without considering these variables, it is difficult to explain behavior. Does a lack of response mean that a user is not interested in the topic, or that that she simply did not see the message?

These probabilistic models also need large amounts of data to learn accurate models, which may not be obtainable for all users.

Most related works are called social targeting or behavioral targeting which learns from past user behaviors, especially feedbacks (i.e., comments, clicks) to match the best advertisements to users.This has resulted in a spike of interest in user data analysis and profile generation as published in [12,31]. In [2], Ahmed et al. presented a time-varying hierarchical user model that captures both the users long term and short term interests. A dynamic topic model was employed. They also showed a streaming distributed inference algorithm that both scales to tens of millions of users and adapts the inferred users interest as it gets to know more about the user. Our work is different in that we focus on collective inference of social brand reputation based on large scale user behaviors.[81]

In [22], authors proposed that Twitter users can be usefully modeled by the tweets and relationships of their Twitter social graphs. They developed a recommender system called Twittomender which demonstrated how user profiles can be used as the basis for recommendation.

Given the widespread generation and consumption of content, it is natural to target ones messages to highly connected people who will propagate them further in the social network.One aspect is the popularity and status of members of these social networks, which is measured by the attention they receive from the consumers of their content. The other aspect is the influence that these individuals wield, which is determined by the actual propagation of their content through the network. This influence is determined by many factors, such as the novelty and resonance of their messages with those of their followers and the quality and frequency of the content they generate.Each user inde- pendently decides what other users to follow. Each retweet explicitly credits the author of the original tweet [6] and is an important influence signal.[61]

Retweets allow one to track the flow of information in Twitter because they indicate situations where a user felt a tweet was important enough that he or she shared it with his or her followers. For this reason, to predict information spreading in Twitter, we wish to predict retweets.

The binary feedback is 1 if the retweeter retweeted the tweet within a certain time window, and 0 otherwise. For our training data, we used a time window of one hour. This is sufficient because it was found in [32] that half of the retweets occur within an hour of the source tweet.

The origin of information propagation in social networks comes from social influence, which occurs when an individuals thoughts, feelings or actions are affected by other people. Information propagation characterizes the way that a node in social networks can spread an information meme to its neighbor nodes via exerting social influence on them. Research from cognitive science has also provided some initial evidence of quantum-like cognitive interference in human decision-making.Twitter users send and read messages called tweets, which contain no more than 140 characters. One user can read another users messages by following them. In addition, one users message can be re-sent by his followers via retweeting. A retweeting message starts with the identifier RT @username. Such following/follower relationships connect Twitter users and form the social network where information flows through retweeting. Given a collection of tweets C = t, V represens all Twitter users while E = (u, v) —u, v V represents all following relations where u follows v.

Some users are prone to elongating words through character repetition, e.g., by writing cooool instead of cool. Brody & Diakopoulos find that this phe- nomenon is common on Twitter and that subjective terms in particular are lengthened in this way, presumably to inten- sify the expressed sentiment [7]. Word elongation may thus be indicative of emotivity, which we hypothesised could be linked to high popularity or influ- ence. Elongating words can also be taken as an indication of a lack of formality.We thus record the fraction of words that a user elongates in this way. Since three or more identi- cal, consecutive letters are very unusual in English, a word is considered elongated if the same character is repeated consecutively at least three times.Twitter is a breeding ground for idiosyncratic uses of lan- guage since the 140 character limit on messages forces users to find new ways of expressing themselves. Aside from the sociolinguistic interest of such a study, there are also practical uses for the identification of powerful or influential actors online, for example, for social media marketing or political campaigning. In both of these areas, since we are increasingly using the Internet as a source of news and opinions, it would be helpful to learn how to be more influential online.Sociolinguistic studies provide the basis for the underlying assumption of this study, namely that individuals with higher social power differ from low status individuals in their use of language. High status users express more emotions than low status users.[73]

In this paper, we present a recommendation algorithm aiming to identify potentially interesting users to follow in the Twitter network.Recent research efforts on understanding micro-blogging as a novel form of communication [][Java et al., 2007; Kr- ishnamurthy et al., 2008] revealed that few users in Twitter maintain reciprocal relationships with other users. This fact differentiates Twitter from other online social networks, such as Facebook, Hi5, or Orkut, in which people mainly make connections to keep in touch with people they consider as friends or acquaintances. Although posts in Twitter or tweets are allowed to have any textual content within the limit of 140 characters, many users only pub-

lish information about a particular subject, such as sports, movies, music or about a particular rock band.[3]

Kwak et al. [32] quantified these findings indicating that 77.9% of Twitter connections are unidirec- tional and only 22.1% of the relations are reciprocate. More- over, 67.6% of users are not followed by any of their followees, indicating that these users probably use Twitter as a source of information rather than as a social networking site.

Chen et al. [69] compared relationship-based and content-based algorithms in making people recommen-dations, finding that the first ones are better at finding known contacts whereas the second ones are stronger at discovering new friends.

Intuitively, a weibo is useful to a user, if the user is interested in or willing to read the weibo. Whether a user is interested in a tweet is determined by many factors, such as the quality of the weibo, the influential degree of the authors, etc. Personal interest is also an important factor to decide whether a tweet is personally useful.

As Twitter has become a popular social medium and had great impact, plenty of researches focus on analyzing the personal interest of users and building recommendation algorithms. Michelson et al [45] detect the entities of each tweet, and discover the topics of interests for Twitter users. Ramage et al [56] applied Labeled topic models to analyze the content of each tweets. Yang et al [79] established a joint friendship- interest propagation model to make link prediction and tweet recommendation in a unified framework. Chen, et al [11] proposed a collaborative personalized tweet recommendation algorithm, which integrate Twitter factors into a unified framework. Yan, et al [78] made Tweet Recommendation by combining three graphs together. Those researches did not take deep insight into how social influence is generated according to users historical record and how the influence will determine the results of tweet recommendation. In this paper, we combined both global Tencent features and topic level social influence into a unified framework, which is proved to gain a better performance than traditional method.

we compartmentalize, or group, users by their activity, interest in the topic and number of friends. This means that the model treats users with the same number of friends and same activity level as indistinguishable from each other when computing transitions between states, thereby ignoring additional individual differences. This simplification makes stochastic models tractable by reducing the number of parameters necessary to describe the population.

Twitter is a popular micro-blogging social media platform that enables communication between networked users. Users (Twitterers) can broadcast an unlimited amount of messages (tweets) to a group of other Twitterers who have opted to subscribe to these broadcasts (followers). Twitterers also receive broadcasts from other users to whose account streams they subscribe or are following. Individual tweets are limited to 140 characters. The Twitter platform supports both broadcasting and receiving tweets through an online web portal and via the text- messaging feature on most mobile phones.

Over the past forty years, social network theorists have been interested in studying the factors at play behind formation of ties between individuals. A tie, typically reflecting a psychological or sociological relationship, defines the topological characteristics of a network at large. It is responsible for affecting the spatio-temporal dynamics encompassing a number of social phenomena in networksincluding how information propagates from one individual to another, or how communities emerge and evolve around shared relationships.

In particular, extracting users interests with the content of document simultaneously has received considerable attention in recent years. This has led to the proposal of various topic models that can infer latent topics with authors interests automatically, where each topic is a latent variable that has a probability distribution over words. These topic models introduce authors interests as topic distributions according to their document contents. These proposed topic models aim at using topics or sets of topics to represent authors interests and provide useful description for the generative process of various data, which could be applied in different aspects such as social network analysis, expertise recommendation and collaborative filtering , etc. [82] However, most topic models combined with authors information have not considered authors sentiment with his interested topics. However, these works could not study well on users interest level but just in documents level.It is essential to identify users sentiment to his interests. According to the previous probability generative model, topics or users interests are extracted only with the probability of co-occurrence, which means if a user talks about one topic frequently, the models will consider that the user is interested in it but do not care the sentiment trend on the topic. Users who talk much about a topic could have different opinions about it. It is better to distinguish these users instead of viewing them in the same community. For example some people talk a lot about the topic they like, while even if others dislike it, they could also discuss a lot about it in negative aspects. In conventional methods, these users will be clustered in the same community, but it is obvious that people with different opinions should be separated.One thing should be noted that our model considers sentiment and topic simultaneously, by which it makes that not all extracted topics have both positive and negative sentiment trends.

2 Related Works

Recently, the researching works about topic model have focused more on two aspects, i.e., topic models combined with

authors information and topic models combined with sentiment information. Topic model is a stream of new research in machine learning and natural language models for clustering words in order to find the underlying topics. Latent Dirichlet Allocation (LDA) [5] can robustly discover multinomial word distributions of these topics. When merging authors information in topic model, some extended works based on LDA have been proposed, Author-Topic model [65] learns topics conditioned on the mixture of authors that composed a document which each author has a distribution over topics, Author-Persona Topic model [46] allows each authors documents to be divided into several clusters in order to find personas of the author by the topic distribution over these documents clusters. Compared with these works that represent authors interests or personas by topic distribution, Author Interest Topic model (AIT) [29] considers adding a document class layer between the author layer and the topic layers which represents the authors interests as a mixture of document classes. Latent Interest Topic model (LIT) [30] is extended from AIT by adding an author class layer to model the relations between different authors, while Role-Author-Recipient model [42] considers the roles of users by assigning a role-layer to the pair of author and recipient. Moreover, researching works have tried to utilize the semantic analysis of topic model with social graph to detect communities. In [83], the authors proposed the CUT (Community-User-Topic) model which discovered communities using semantic content of the social graph. One of the first attempts to combine link based community discovery methods with content based method is the Community-Author- Recipient Topic (CART) model [50]. However, since CART, there have been few advanced attempts on combining link and content analysis to detect communities more effectively. The Topic-Link LDA model [38] and Rational Topic Models [9] draw latent topical and community distributions for each node in documents networks. Then they generate links between documents based on topical similarity and community membership similarities of their authors. All these works have shown the availability of topic models to do Social Network Analysis but they all ignore the sentiment information among the topics. When combined sentiment detection with topic models, this probability generative model works also show a strong suitability. Several unified models of topics and sentiment have been proposed, and they extend basic topic model works to explain the sentiment trends with topic from documents such as reviews or comments [44,37,27]. Topic Sentiment Mixture (TSM) model [44] represents the sentiment as a language model separated from topics, which means TSM considers the topic and sentiment separately, the word samples from either topics or sentiments. Multi-Aspect Sentiment (MAS) model [27] 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 [37] 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. JST takes much similar way as our work but it only detects the sentiment on document level. Moreover, to limit the meaning of extracted topics, extra labels or constraints are incorporated into topic model frameworks. Labeled-LDA [57] is a supervised topic model for credit attribution in multi-labeled corpora, which constrains each topic from particular labels by assigning different prior probabilities from multi-labels in the corpora. From this definition, the topics extracted by Labeled-LDA are more meaningful and distinguishable towards labels. Our proposed model learns the idea from Labeled-LDA to distinguish the sentiment labels of topics. However, our model needs only a general sentiment paradigm word list but Labeled-LDA needs to provide labels to each document in the corpora.

To the best of our knowledge, such an integration problem has not been studied in the existing work.[39]

Furthermore, a lot of current opinion mining work focuses on mining review data and solving classification problems. As we go beyond product reviews, only knowing sentiment orientations such as positive, negative and neutral is not enough in many cases. This is especially true in the domain of politics where the wording is often sensitive. For exam- ple, with respect to healthcare reform in U.S., a Republican might often say we want responsible healthcare reform based on private insurance1, while a Democrat might often say we want universal healthcare reform with a public government-run health insurance agency. Both statements can be viewed as positive on healthcare reform in general, but the opinion wordsresponsibleandprivatevsuniversal and public reflect their huge difference on the issue.Therefore, in COM, the opinions of interest are represented by opinion words which are directly returned to users.[18]

As the reason of that, Twitter has three features: convenience,immediacy and propagating.[25] Twitter facilitates real-time propaga- tion of information to a large group of users. This makes it an ideal environment for the dissemination of breaking-news directly from the news source and/or geographical location of events.[8]

Review data is relatively easier to work with since sentences in reviews are more grammatically correct compared to tweets. Tweets on the other hand are short, informal, colloquial, and can contain slangs and abbreviations. Thus, many of the existing natural language processing tools such as Part-of-Speech (POS) tagger, tokenizer, and dependency parser fail to work well because they are typically trained on newswire data.Instead, current Twitter sentiment analysis approaches

[4,16,21,58,62] adopt a machine learning approach similar to text classification.Training data are often obtained by using noisy labels (also known as distant supervision). This allows us to classify the tweets without manual supervision. Go et al. [21] and Read [58] exploit emoticons such as :-) and :-(to label the tweets positive and negative respectively. They then treat the problem as a text classification task and use machine learning techniques such as Nave Bayes (NB), Maximum Entropy (MaxEnt), and Support Vector Machines (SVM) to train a classifier. Barbosa and Feng [4] on the other hand, construct their training data from a few different Twitter sentiment analysis websites.[34]

For example, there are several words that are written informally like nooooooooooooo, loveeeeeeeeeee,sundayssss, saddddddd,xxplosive, fooooooooood,okaaay and so on. A sentiment is often represented in subtle or complex ways in a text. An online user can use a diverse range of other techniques to express his or her emotions. Apart from that, s/he may mix objective and subjective information about a certain topic. On top of that, data gathered from the World Wide Web often contain a lot of noise. Indeed, the task of automatic sentiment recognition in online text becomes more difficult for all the aforementioned reasons.

For example, Tan et. al. [71] utilized the social connection to improve the sentiment classification performance based on the intuition that connected users are more likely to share similar opinions towards the same entity.

in many applications, the polarity should be on the user instead of a single document. In fact, the quality of a single document (e.g., a product review) may vary largely, even for the same user. [72]

Tan et al. [71] study the problem of user-level sentiment analysis using social networks. we employ the assumption that the influence between the nodes only occurs within distance of 1. Thus a users sentiment is only influenced by himself and his followees.We work within a semi-supervised, user-level framework. The reason we adopt a semi-supervised approach is that the acquisition of a large quantity of relevant sentiment-labeled data can be a time- consuming and error-prone process, as discussed later in this pa- per. We focus on user-level rather than tweet-level (corresponding to document- or sentence-level) sentiment because the end goal for many users of opinion-mining technologies is to find out what peo- ple think; determining the sentiment expressed in individual texts is usually a subtask of or proxy for that ultimate objective. Addition- ally, it is plausible that there are cases where some of a users tweets are genuinely ambiguous (perhaps because they are very short), but his/her overall opinion can be determined by looking at his/her collection of tweets and who he/she is connected to. when a user forms a link in a net- work such as Twitter, they do so to create a connection. If this connection corresponds to a personal relationship, then the prin- ciple of homophily [33] the idea that similarity and connection tend to co-occur, or birds of a feather flock together [43]sug- gests that users that are connected by a mutual personal relation- ship may tend to hold similar opinions; indeed, one study found some evidence of homophily for both positive and negative sentiment among MySpace Friends [75]. Alternatively, the connection a user creates may correspond to approval (e.g., of a famous figure) or a desire to pay attention (e.g., to a news source), rather than nec- essarily a personal relationship; but such connections are still also suggestive of the possibility of a shared opinion. First, we empirically confirm that the probability that two users share the same opinion is indeed correlated with whether they are connected in the social network. since the Twitter interface makes the tweets of t-followee vj visible to t-follower vi (and similarly for @-mentions), so we have some reason to believe that vi is aware of vjs opinions. In this section, we frame the problem in the context of Twitter to keep things concrete, although adaptation of this framework to other social-network settings is straightforward.[71]

Emotion is now recognised as an important aspect of many areas of our lives. Aside from the obvious relevance of feelings like happiness and sadness to personal wellbeing, appropriate perception and communication of emotion is important for maintaining human relationships and friendships and not just intimate relationships. Particularly for women, friends are also a source of emotional support, one that can help individuals to cope in difficult times.[67], it is important to understand as much as possible about the properties of online emotion expression and support in friendship, so that appropriate actions can be taken if the important social institution of friendship is under threat or, conversely, if new opportunities are evident.

There are many socially recognised types of emotion, such as happiness, fear,sadness and anger, but one common way of analysing emotion is in terms of the two dimensions of valence and arousal [15]. Valence is the type of emotion felt: the extent to which it is positive or negative. Arousal is the general perceived level of activation or energy and it is often associated with the strength of an emotion. Hence it is reasonable to simplify a discussion of emotions in friendship to just valence/polarity and strength. Note, however, that it is possible to simultaneously experience or express positive and negative emotions [20]: to have mixed feelings about an event. An emotion categorisation program (Thelwall, et al., submitted) was used to assign a positive and a negative emotion to each comment on a scale of one (no emotion) to five (very strong emotion). The classification was based upon a predefined list of about 500 emotionbearing words (e.g., hate = negative 4), emoticons and emphatic devices including words (e.g., very), punctuation (e.g., !!!!) and repeated letter emphatic spelling (e.g., soooo).[75]

The central issue addressed in this paper is homophily, the tendency for friendships and many other interpersonal relationships to occur between similar people. Based upon a survey of predominantly US research, it seems that gender, sexuality, religion, race and age similarity are all important predictors of friendship[43].[74]

micro-blogs differ by (1) placing a strict limit on length, resulting radically in new forms of emotional expression, and (2) encouraging users to express their daily thoughts in real-time, often resulting in far more emotion statements than might normally occur. Micro-blogging services such as Twitter provide researchers with a wealth of information on how individu- als communicate with their social network. Unlike more formalmethods of communication,micro-blog posts (here- after, tweets) frequently reflect the authors opinions and emotional states. For instance, Table 1 shows several re- cent tweets reflecting on the latest FIFA World Cup. Fur- thermore, since tweets are restricted to 140 characters, and since they are oftenwritten onmobile devices, they express emotions less formally than other publishing platforms.Such a systemwould be useful in understanding users feelings towards particular products, services, or topics (e.g., companies could deter- mine the distribution of emotions toward their latest product).[60]

A prerequisite of all such research is an effective method for measuring the sentiment of a post or tweet. Due to the extremely informal nature of the medium, and the length restriction1, the lan- guage and jargon which is used in Twitter varies sig- nificantly from that of commonly studied text cor- pora. In addition, Twitter is a quickly evolving domain, and new terms are constantly being introduced. In purely text-based domains, such as Twitter, styling is not always available, and is replaced by capitalization or other conventions (e.g., enclosing the word in as- terisks). Additionally, the informal nature of the do- main leads to an orthographic style which is much closer to the spoken form than in other, more formal, domains.the commonly observed phenomenon of lengthening words by repeating letters is a substitute for prosodic em- phasis (increased duration or change of pitch). As such, it can be used as an indicator of important words and, in particular, ones that bear strong in- dication of sentiment.Phenomena include mis- spellings, abbreviations (e.g. gr8 - great), emphatic uppercasing (WHAT THE HELLWAS THAT????), emphatic lengthening (The concert was greeeeeeeeeaat!!!) and the use of slang and neologisms. This leads to much more sparsity in the input and is a special chal- lenge for the use of lexical resources. Brody and Diakopoulos [7] find emphatic lengthening to occur in every 6th tweet of their dataset and provide a detailed analysis.[7]

Beside being an interesting research problem, sentiment analysis can be di- rectly applied by persons interested in a large amount of opinions towards a certain topic. With the rise of social media, the number of opinions on the web has multiplied, as platforms like Facebook and Twitter make it very easy for ev- eryone to share their thoughts on literally anything. This calls for sentiment analysis systems that can process large amounts of data and are able to han- dle the special challenges of the text genre of so-called microblogs. Because of the interest in utilizing this freely available information by research and industry, sentiment analysis of microblogs has become a popular research topic during the last years. Besides the challenges traditional sentiment analysis systems face, such as ambiguity, handling of negation, detection of sarcasm and opinion spam, sentiment analysis of microblogs have to handle the following additional difficultiesText Length: Microblog posts are usually very short. While this can be an advantage, because authors tend to get straight to the point they want to make, it poses the challenge that the expressed opinion might be dependent on one word only. The word might not be available in the used lexical resource or might not have occurred in the training data, which can lead to the loss of the opinion. A discussion of the phenomenon can be found in Bermingham and Smeaton (2010). Spelling variation: Due to spontaneity, the informal context and length restrictions the spelling in microblog posts tends to have much greater variability than in other text genres. Phenomena include mis- spellings, abbreviations (e.g. gr8 - great), emphatic uppercasing (WHAT THE HELLWAS THAT????), emphatic lengthening (The concert was greeeeeeeeeaat!!!) and the use of slang and neologisms. This leads to much more sparsity in the input and is a special chal- lenge for the use of lexical resources. Brody and Diakopoulos (2011) find emphatic lengthening to occur in every 6th tweet of their dataset and provide a detailed analysis. Special tokens: Tokens uncommon in other text genres, such as URLs and emoticons, can lead to difficulties when trying to use natural language processing tools, such as part-of-speech taggers and syntactical parsers. The latter are often trained on newspaper texts, which are considerably different to microblog posts. Topic variation: The topics discussed on Twitter are not constrained in any way and the variety is therefore very large. This can cause problems for sentiment analysis, e.g. when words express different sentiment in different contexts. Amount of data: While the texts as such are often short, the amount of texts can be overwhelmingly large. In 2012 the popular microblog- ging service Twitter announced1 12,233 posts per second about the American football Super Bowl towards the end of the game. Language style: Due to Twitters large userbase the variety in writing style is very large. This might range from formal newspaper-like text to very informal slang including profanity. Furthermore, the vocabulary used can change rapidly. All this can lead to problems for annotated training data and lexical resources. Multilingual content: While online newspapers and blogs tend to be written in one language, users of microblogging platforms use a wide vari-

ety of languages, sometimes even in the same message or sentence. With the shortness of the posts language detection becomes increas- ingly difficult. These difficulties do not apply to sentiment analysis exclusively, but are also of concern for other natural language processing tools, such as part-of-speech taggers, parsers and the like.

Therere two important factors that should be taken into consider- ations. One, opinions and topics are closely related.The online discussions around some entity, or object, often cover a mixture of features/topics related to that entity with different preferentials. Different opinions may be expressed by users towards different topics, where users may like some aspects of an entity but dislike other aspects. Two, users opinions are subject to social influence. The rise of social media puts the sentiment analysis in the context of social network. Users not only express their individual opinions, but also exchange opinions with others. In the context of opinion mining, social influence refers to the phenomenon that one is inclined to agree (positive influence) or disagree (negative influence) with his/her neighbors opinions with different degrees, depending on the influence strengths.[35]

As a real-time conversational platform Twitter presents a number of interest-ing user modeling and profiling opportunities and challenges. Twitter users are connected in relatively dense social networks of followers and friends and these networks can contain communities of users with shared interests. In addition, the conversations that take place within these networks, and the information that is shared through these conversations, has the potential to provide a rich source of user preference and interest data, notwithstanding the 140-character limit that is placed on user posts.In recent work [22], the text content of Twitter messages from a user and their followers/friends was used as the basis for text-based profiles as part of a recommendation system to suggest new users to follow. In this work, recommended users were suggested on the basis of term overlaps between the target users profile and the profiles of other users. Other researchers have ex- plored Twitter information to model users as the basis for news recommendation [**?**,54] based on term overlap between news story content and user tweets. One of the challenges with using tweet content as the basis for profiling is that it can lead to very large but noisy user profile [36].We propose that this regioned, multi-faceted approach to profiling Twitter users provides an effective way to model the interests and preferences of users in a way that facilitates a better understanding of core and peripheral interests and also provides for a powerful framework for exploring a space of interests within Twitters social graph.

Twitter is well-known for its freedom of publishing short message (i.e. tweet), and viral spreading of information across complex social networks. In addition to large amounts of User-Generated Content (UGC), Twitter provides its social network functions for connection, communication and information diffusion by allowing users to message one another directly and follow one another publicly. The complex networks and large content volume of Twitter provide researchers with insights into peoples social behaviors on a scale that has never been possible [66].

Information diffusion is a challenging problem which might be investigated on Twitter, because retweeting convention and complex networks of Twitter have provided an unprecedented mechanism for the spread of information despite the restricted length of tweets [26]. Actually almost 25% of the tweets published by users are retweeted from others [80]. Therefore, it is important to understand how retweeting behavior works which can help study information diffusion on Twitter.

Although several works have concentrated on analyzing retweeting habits and influencing factors [6,32,68], most of them are generic, not user-oriented. From the point of a user, retweeting is a process that includes reading the tweet, estimating the content and deciding to share, and the crucial part of the process is to estimate whether a tweet contains information interesting to the user who might find it worthy to be shared. Therefore in this study we focus specifically on analyzing the retweeting behavior from the user modelling perspective.

Previous studies on retweeting analysis have shown that an enriched user model gives coherent and consistent explanation for retweeting motivation[**?**,41,19]. Specifically, researchers have tried to model users from four types of information: profile features ("**Who you are**"), tweeting behavior ("**How you tweet**"), linguistic content ("**What you tweet**") and social network ("**Who you tweet**") [**?**]. Despite demographic profile, tweeting habits and network structure might determine the source and scope of information users could be exposed to, topics of interest encapsulated in rich linguistic content have been proved consistently dependable for retweeting behavior explanation. For example, Petrovic *et al.* [52] and Hong *et al.* [23] found whether a tweet will be propagated largely depends on its identification with the interests of users. However, beyond merely publishing news and events, Twitter has become a platform where different opinions are presented and exchanged by allowing users publish subjective messages on topics they are interested in. Existing researches demonstrated that UGC with rich sentimental information can trigger more attention, feedback or participation [66], and tweets with high emotional diversity have a better chance of being retweeted [53]. Most studies have tried to find whether and how sentiment of a tweet will influence its spreading, but none of them realize that although users receive thousands of tweets on different topics every day, whether a tweet will be retweeted will depend on the subjective choice of users.

Subjective initiative nature of human determines that his behavior pattern is subjectivity driven. Psychologist have

identified subjectivity as the underlying factor that influences taking what behaviors to process incoming stimul [47]. According to theory of Biased Assimilation, people are prone to choose and diffuse information according to their own biased subjectivity [24,70]. In this study we explore the UGC of Twitter to model the subjectivity of users, and investigate whether the subjectivity model could benefit the retweeting behavior analysis. Intuitively, subjectivity can be represented as topics and opinions articulated in the information generated by users on Twitter. We use the state-of-the-art topic model to find the topics users are talking about, and sentiment analysis techniques to determine user's opinions towards these topics from UGC simultaneously. We evaluate our model on the retweeting analysis problem to verify its impact on retweeting behavior.

Modelling subjectivity on Twitter is a challenging task because of the sparsity of textual information and the dynamic of topics and opinions. However, we are interested in understanding retweeting behavior at a local level rather than at a global level, since most of time retweeting pertains to a local network consisting of the tweet publisher and followers, and the relatively tiny size and topic homophily of local network 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 is likely to be more consistent and less prone to external perturbations, therefore subjectivity model of a user is less likely to be influenced by changes of topics and opinions on Twitter.

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 for users and tweets which model both the topics and opinions simultaneously.
- Based on the state-of-the-art topic model and sentiment analysis techniques, we build subjectivity model from UGC on Twitter and apply it to the retweeting behavior analysis problem.
- We systematically evaluate the effectiveness of the subjectivity model. It is demonstrated that our model outperforms other UGC-based models in rewteeting prediction and gives the most significant improvement over a off-the-shelf predicting model.

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

3 Related Work

Retweeting Analysis. A lot of works have analyzed the characteristics of retweeting, examining factors that lead to increased retweetability and designing models to estimate the probability of being retweeted. As for factors influencing retweetability, Suh *et al.* [68] found that tweets with URLs and hashtags were more likely to be retweeted. Macskassy and Michelson [41] found that models derived from tweet content could explain most of retweeting behaviors. Comarela *et al.* [14] 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 [63] found that tweets with topical keywords were more likely to be retweeted. There are also many works extending the analysis to build retweeting prediction model. Osborne and Lavrenko [52] introduced features such as novelty of a tweet and the number of times the author is listed to train a model with a passive aggressive algorithm, and found that tweet features added a substantial boost to the performance. Jenders *et al.* [26] analyzed the "obvious" and "latent" features from structural, content-based, and sentimental aspects and found a combination of features covering all aspects was the key to high prediction quality. Naveed *et al.* [49,48] introduced interestingness based on such features as sentiments and topics to predict the probability of retweeting for an individual tweet. Feng and Wang [19] proposed a feature-aware factorization model to rerank the tweets according to their probability of being retweeted. Pfitzner *et al.* [53] proposed a new measure called emotional divergence and showed that high emotional diverse tweets have higher chances of being retweeted.

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 ". In this paper, we will try to answer this question by building a subjectivity model which can capture both the interests and opinions of users.

User Modelling. With the popularity of social media, researchers have begun to pay close attention to model users on the massive amount of UGC. These studies provide researchers with insights into user online behaviors. Hannon *et al.* [**?**] proposed that Twitter users can be modeled by tweets content and the relation of Twitter social network. Macskassy and Michelson [41] discovered user's interests by leveraging Wikipedia as external knowledge to determine a common set of high-level categories that covers entities in UGC. Ramage *et al.* [**?**] made use of topic models to analyze tweets at the level of individual users with 4S dimensions, showing improved performance on tasks such as post filtering and user recommendation. Xu *et al.* [**?**] proposed

a mixture model which incorporated three important factors, namely breaking news, friends' timeline and user interest, to explain user posting behavior. Pennacchiotti and Popescu [**?**] proposed a comprehensive method to model users for user classification, and confirmed the value of in-depth features by exploiting the UGC, which reflect a deeper understanding of the Twitter user and the user network structure.

Few of work have identified the correlation between the opinions of users and their behaviors, motivated by the observation, we put forward subjectivity model to combine both interests and opinions to model a user.

Sentiment Analysis. Sentiment analysis is a popular research area and previous researches have mainly focused on reviews or news comments. Recently, researchers began to pay more and more attention to social media such as Twitter. Hu *et al.* [**?**] interpreted emotional signals available in tweets 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.* [**?**] focused on target-dependent Twitter sentiment classification, and proposed a method to improve performance by taking target-dependent features and related tweets into consideration. Asiaee T. *et al.* [**?**] 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. [**?**] 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. Motivated by sociological theories that humans tend to have consistently biased opinions, Calais Guerra *et al.* [**?**] 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.* [**?**,**?**] 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 subjectivity model , because the fine-grain sentiment strength it outputs could give us more detailed opinion than binary labels.

4 Subjectivity Model

In this section, we firstly give the definition of subjectivity model, then describe the method of building subjectivity model, and finally apply subjectivity model to the retweeting analysis problem.

4.1 Definition

Subjectivity has been extensively studied by psychologists to characterize the personality of a person based on his historic behaviors and remarks [17]. Linguists define the subjectivity of language as the speakers always show their perspectives, attitudes and sentiments in their discourses [64]. Social media provides users a platform to express their opinions towards topics of interest to show their personal subjectivity by publishing short messages. Therefore, for the term "subjectivity" , we refer to both topics and opinions articulated in the UGC. That is, we model subjectivity not only by interests of users, but also by "**what they think about the interests**". Here we firstly give our definition of subjectivity model on Twitter, while we emphasize that our model can be transfered to other social media platforms as well.

For a set of users U on Twitter, we assume there is a topic space T containing all topics they talk about, and a sentiment valence space S evaluating their opinions towards these topics. As for S, it is often considered as a binary space consisted of positive and negative sentimental values, however we argue that a more fine-grained sentiment space will indicate more detailed opinions of users.

Definition 1 (Subjectivity Model For User) The subjectivity model $P(u)$ of a user $u \in U$ is the combination of a set of topics $\{t_i\}$ the user talks about in a topic space T and the user's opinions $\{O_i\}$ towards the topics.

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

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, subject to $\sum_{i=1}^{|T|} w_u(t_i) = 1$.
- opinion O_i of user towards topic t_i is a target-dependent sentiment distribution $d_{u,t_i}(s_i)$ over sentiment valence space S, subject to $\sum_{i=1}^{|S|} d_{u,t_i}(s_i) = 1$.

Users express themselves by tweeting on Twitter, and each tweet generated by users can be considered subjective in that it also contains topics and opinions. So we also give a subjectivity model definition for a tweet as follows:

Definition 2 (Subjectivity Model For Tweet) The subjectivity model $P(m)$ of a tweet m is the combination of a set of topics $\{t_i\}$ it talks about, and the opinions $\{o_i\}$ it expresses.

$$P(m) = \{(t_i, w_m(t_i), \{d_{m,t_i}(s_i))\}, |t_i \in T, , s_i \in S\} \quad (2)$$

The definition of subjectivity model given above is in an abstract form by using latent concepts of topics and opinions which need to be derived from UGC. In this paper we combine subjectivity model with retweeting analysis problem and concrete the subjectivity model in such problem settings.

4.2 Retweeting Analysis Problem Statement

Retweeting is the core mechanism of information diffusion on Twitter. Many factors have been proved to influence retweeting behavior [68,41,14], however few researches have investigated the subjective motivation of a user to retweet a message. Therefore we will study whether subjectivity model can help understanding underlying reasons of a user's retweeting behavior.

In fact the likelihood of a tweet to be retweeted depends on both context constraints and its content. The context such as the network of the author and the time a tweet is published affects whether the tweet will be retweeted. A tweet with only few or passive followers is less likely to be retweeted, and a tweet published in the night have less chance to be retweeted than daytime. Apart from the context constraints, a tweet is more likely to be retweeted by subjective users who find its content worth to. Therefore, we are not interested in modelling the tweet by itself just as other researches [49,48,53], but understanding how the content resonate with the users who might want to retweet it. We put a much stronger emphasis on the content and try to model the user's subjective decision by deriving latent topics and opinions from UGC. Actually, none of contextual factors has any influence on the content of a tweet, therefore we deliberately ignore context constraints to avoid introducing contextual bias into our analysis by proposing Hypothesis 1.

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, the motivation of a user to retweet a message lies in that the user considers only the tweet content arousing his resonance without context perturbation.

On 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 connected in a social network "are homogeneous with regard to many socio-demographic, behavioral, and intra-personal characteristics" [43]. In other words, homophily implies that a user follows another user because he finds they share similar interests. According to the principle of homophily, we put forwards the concept of **Local Topic Space**, which could be defined as:

Definition 3 (Local Topic Space) In a local network consisting of a user and his followers, all users concentrate on limited topics derived from their UGC, and these topics form a local topic space.

Since most of time retweeting pertains to a local network, we limit our research in understanding retweeting behavior at a local level rather than at a global level, and the relatively tiny size and topic homophily of local network lower the impact of data sparsity.

According to our Hypothesis 1, if a tweet is published, all followers of its author will receive it in time, and followers are likely to retweet it if they find it worthwhile. Thus the retweeting analysis problem we study can be stated as follows:

Let F, A, M denote the follower set, author set and tweet set respectively. For each tweet m ($m \in M$) and its listener f ($f \in F$), we can define a quadruple $< f, a, m, r_{fam} >$ where:

- a ($a \in A$) is the author of the tweet m and f ($f \in F$) is a follower of author a.
- r_{fam} is a binary label indicating whether m is retweeted by f.
- Our work focuses on using subjectivity model to analyze the relation between the subjectivity of a follower f and his retweeting behavior. Hence we transform the quadruple into the Local Topic Space T formed by the author a and followers F, and represent f, a, m with their subjectivity models to analyze their relations with the label r_{fam}.

4.3 Establishment of Subjectivity Model

According to the definition of subjectivity model , there are two distributions to model the subjectivity: the topic distribution and the opinion distribution for each topic. Both of them need to be inferred from historic content produced by users. However, content analysis on Twitter is challenging: the volume of tweets is so huge while a single tweet is very short with limit of 140 characters, and informal languages are widely used, which make many supervised learning approaches and natural language processing techniques invalid. Hence effectively modeling content on Twitter requires techniques that can readily adapt to these challenges and require little supervision. With state-of-the-art topic model and sentiment analysis techniques, we establish subjectivity model by identifying topics and opinions in an unsupervised way simultaneously.

4.3.1 Topic Analysis

The topics of a tweet are latent and have to be inferred from its content. Previous studies have tried to identify topics from tweets by finding key words [10], extracting entities [**?**] or linking tweets to external knowledge categories [41], however, the sparsity is a main problem for these methods because even users have common local topics they still might refer to a topic with different vocabulary. Works show that topic models such as **Latent Dirichlet Allocation (LDA)** model and its extensions[5,77] have been efficient ways to characterize latent topics of large volume corpus. Topics of LDA are broader in concept, since a single topic consists of the whole collection of related words. Therefore we adopt a

user-level LDA model to find latent topics for users in their Local Topic Space, and the generative process can be graphically represented using plate notation in Figure 1. To distill the topics that users are interested in, documents of LDA should naturally correspond to tweets content. As our goal is to understand the topics that each user is interested in rather than the topics each single tweet talks about, we aggregate the tweets published by each user into a single document, and replace documents of LDA with aggregated tweet documents. So a document stands for a s user in our model, and a user can be represented as a multinomial distribution over topics, which corresponds to the topic distribution of the user's subjectivity model.

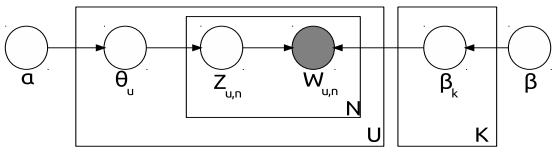

Fig. 1 Plate illustration of the user-level LDA model.

Formally, given a set of users U and the number of topics K, a user u ($u \in U$) could be represented by a multinomial distribution θ_u over topics with a Dirichlet prior parameterized by α. A topic k ($k \in K$) is represented by a multinomial distribution β_k with another Dirichlet prior parameterized by η. 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 [59].

4.3.2 Opinion Analysis

Users often express opinions towards their topics of interest by publishing topic-related tweets. In order to explore the opinions of users, we need to understand sentiment embedded in each tweet. Sentiment analysis mainly depends on machine learning or rule-based approaches. Machine learning approaches often need labelled data for the training process, which is often impossible for Twitter because of the large volume of tweets and its dynamic language characteristics. Therefore we adopt rule-based approaches, which could adapt to Twitter with good flexibility by changing its particular characteristics into rules [**?**,**?**].

The SentiStrength package 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 suitable for analyzing sentiment of tweets in our research settings. SentiStrength assigns two values to each tweet standing for sentiment strengths: a positive (within $[1,5]$) and a negative (within $[-5,-1]$) sentiment measurement, both ranging from 1 to 5 on absolute integer scales, with 1 denoting neutral sentiment and 5 denoting highest sentiment strength. Sentiment assigned by SentiStrength is not a simple binary value but a fine-grained strength, which can catch fine opinion distributions in a user's subjectivity model. For the convenience calculation, we map the output of SentiStrength to a single-scaled sentiment valence space $[0,8]$ as follows:

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

Where p denotes the positive setiment value and n denotes negative sentment value. In the sentiment valence space $[0,8]$, value 4 indicates neutral sentiment, while values above 4 indicate positive sentiment and values below 4 indicate negative sentiment. With the sentiments of all tweets, we can aggregate opinions towards a topic as a sentiment distribution over sentiment valence space $[0,8]$.

4.3.3 Concrete Subjectivity Model

With statistical topic analysis and opinion analysis described above, we can concrete subjectivity model in a local network settings now. For user set U of a local network, we denote tweet set published by a user u as $M_u = \{m_i | i \in [1,\cdots,N]\}$. Each M_u is concatenated to a single document d_u to construct Local Topic Space $T = \{t_i | i = 1,\cdots,K\}$. A topic model is built with parameter θ representing the distribution of each user over topics in the Local Topic Space T, and parameter β represents the distribution of each topic over the vocabulary of all tweets. SentiStrength is applied to each tweet m in collection M_u and outputs sentiment strength s_m for tweet m. We build the subjectivity model $P(u)$ for user u as Algorithm 1:

In the algorithm, we assume the sentiment of tweet m is related to every topic it talks about in Z_m for simplicity. Accordingly subjectivity model $P(m)$ for tweet m as:

$$P(m) = \{(t, p(t|\theta,\beta), d_{m,t}(s)) = 1.0 | t \in Z_m, s \in S\} \tag{4}$$

Noting that, the opinion towards each topic is a distribution of 1.0 on a single sentiment value s of tweet m.

4.4 Retweeting Analysis With Subjectivity Model

To understand the underlying reasons why a user retweet a message, we try to simulate the subjective decision-making procedure by investigating the relationship among the subjectivity models of a tweet, its author and followers. We assume that a user retweet a message because the user not only finds its topics interesting but also shares similar opinions towards these topics. In other words, if the subjectivity models of a tweet and a user are similar enough, the user will

Algorithm 1 Establishment of subjectivity model .

Input:
The user set of a local network, U;
The tweet set published by each user u, M_u;
Output:
The subjectivity model for each user u, $P(u)$;
Topic analysis with a user-level LDA as Section 4.3.1, getting a topic model $P(\theta,\beta|M_u,U)$;
for all tweet $m \in M_u$ **do**
  Sentiment analysis as Section 4.3.2, outputting sentiment of m, s_m;
end for
for user $u \in U$ **do**
  the topic distribution is the corresponding component of parameter θ, θ_u;
  the topics he tweets about are $Z_u = \{t|p(t|\theta_u) > 0, t \in T\}$;
end for
for tweet $m \in M_u$ **do**
  topics of m can be identified by the topic model, $Z_m = \{t|p(t|\theta,\beta,Z_u) > 0, t \in T\}$;
end for
for each topic $t \in Z_u$ **do**
  for sentiment value $s \in S$ **do**
    count the number of tweets which talk about topic t with sentiment value s, $N_s = \sum_{m \in M_u} I(s_m), if s_m = s \& t \in Z_m$;
  end for
  calculating opinion towards topic t, $O_t = \left\{\frac{N_s}{\sum_{s \in S} N_s}\right\}$;
end for
establishing subjectivity model of user u, $P(u) = \left\{\left(t, p(t|\theta_u), \left\{\frac{N_s}{\sum_{s \in S} N_s}\right\}\right) | t \in Z_u, s \in S\right\}$;
return $P(u)$;

have a very high probability to retweet. We call this phenomenon as "resonance", and assume that the resonance between a tweet and users will elicit retweeting behavior. With the subjectivity models built for users and tweets, we can define a similarity measurement to quantify the resonance among them.

Formally, for a tweet m, the corresponding author a, and a list of followers $F = \{f_i\}$, for each $f_i \in F$, we can define a quadruple $< f_i, a, m, r_{fam} >$ as Section 4.2. We firstly build subjectivity model $P(u)$ for each user $u \in F \bigcup p$ and $P(m)$ for tweet cm, then define the similarity measurement as follows:

$$Sim(m, f_i) = similar(P(m), P(f_i)) \tag{5}$$

according to Equation 18,4:

$$Sim(m, f_i) = \lambda * Dist\left(p(t|\theta,\beta,Z_m), p\left(t|\theta_{f_i}, Z_{f_i}\right)\right) + (1-\lambda) * \left(\sum_{t \in T} Dist\left(O_{m,t}, O_{f_i,t}\right)\right) \tag{6}$$

where

- λ is the coefficient used to control the proportions of topic similarity and opinion similarity in the holistic subjective similarity. We initiate it by setting $\lambda = 0.5$, and adjust its value in the range of $[0,1]$ to optimize the best performance of our model.

Table 1 Retweet Dataset Statistics

Tatget tweets	500
Average number of followers per target tweet	89
Total retweeters	5214
Total non-retweeters	40317

- $Dist()$ is the similarity measurement between two distribution, we choose *cosine distance* in our research among different measurements because of its effectiveness.

We also assume that a user might retweet another user because of their subjective resonance. Accordingly we define similarity between author a and follower f_i as:

$$Sim(a, f_i) = \lambda * Dist\left(p(t|\theta_a, Z_a), p\left(t|\theta_{f_i}, Z_{f_i}\right)\right) + (1-\lambda) * \left(\sum_{t \in T} Dist\left(O_{a,t}, Of_i, t\right)\right) \tag{7}$$

5 Experiment

In this section, we investigate whether subjectivity model can help retweeting analysis on an Twitter dataset.

5.1 Dataset

We adopt an off-the-shelf Twitter dataset of previous work [40], which was created with Twitter API[1]. To form the dataset, 500 randomly selected English target tweets were monitored in the next few days to find followers who would retweet them. Also each target tweet was chosen as starting point to collect historical data of its author and followers. Overall, there are 45,531 followers and 6,277,736 tweets, and 5214 followers who have retweeted at least one target tweet during the monitored period. Summary statistics of the dataset are listed in Table 1.

The relations among a target tweet, its author and followers are illustrated in Figure 2. There is a local network

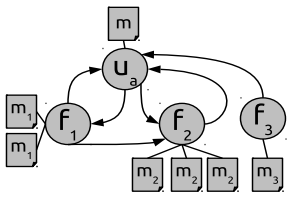

Fig. 2 Relations among a target tweet, its author and followers.

structure for each target tweet as figure shows, consisting of its author and followers.

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

5.2 Impact Evaluation of Different Factors

In Section 4.4, we model retweeting probability with subjectivity model in the form of similarity measurements 6,7. By setting different value to λ, the measurements can be transformed into different versions to model different factors that might influence user's retweeting behavior, which are:

- **TTF**: **T**opic similarity between **T**weet and **F**ollower ($\lambda = 1$ in measurement 6).
- **OTF**: **O**pinion similarity between **T**weet and **F**ollower ($\lambda = 0$ in measurement 6).
- **STF**: **S**ubjective similarity between **T**weet and **F**ollower ($\lambda \in (0,1)$ in measurement 6).
- **TAF**: **T**opic similarity between **P**ublisher and **F**ollower ($\lambda = 1$ in measurement 7).
- **OAF**: **O**pinion similarity between **P**ublisher and **F**ollower ($\lambda = 0$ in measurement 7).
- **SAF**: **S**ubjective similarity between **P**ublisher and **F**ollower ($\lambda \in (0,1)$ in measurement 7).

The six similarity measurements could be grouped into two aspects. One is consisted of TTF, OTF and STF, which is direct and explicit by modelling the tweet and its followers; the other is consisted of TAF, OAF and SAF, which is indirect and implicit by modelling the author and follower. The two aspects reflect properly the local information diffusion structure of Twitter at micro-level as illustrated in Figure 2.

To evaluate the impact of different factors on retweeting behavior, we compare six average similarity scores between 5214 retweeters and 5214 randomly selected non-retweeters. The values of λ for STF and SAF are tuned to produce the largest value difference between retweeters and non-retweeters, which are $\lambda = 0.5$ on our dataset. Figure 3 shows the result. As the figure illustrated, the similarities scores of retweeters are obviously higher than non-retweeters for all six factors. Specifically:

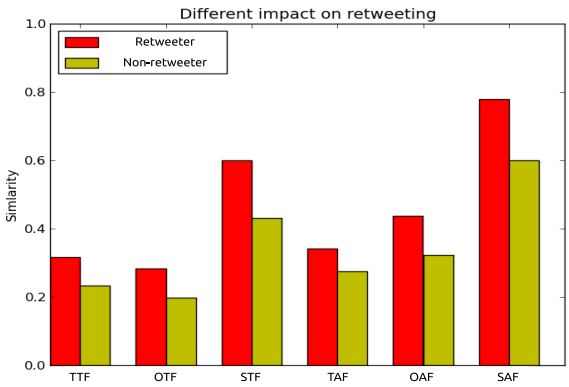

Fig. 3 Impact of different factors on retweeting behavior.

- 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[41];
- OTF score shows that opinions in a tweet is an important indicator to be retweeted by by followers who hold similar opinions, although other studies[53,48] have shown that sentiment in tweet has impact on retweeting behavior, they haven't consider the opinions of followers and opinion similarity betweet tweet and followers;
- STF score shows the subjective similarity is the most distinguishable feature among the six factors with the largest value difference, which proves the importance of subjectivity model;
- TAF score gives another perspective for retweeting analysis from the topic similarity between author and followers, indicating that followers are more likely to retweet author with similar interests , which verifies the homophily principle of following relation;
- OAF score indicates that similar opinions also influence followers' decision of retweeting another user, which proves opinion homophily of following relation.
- SAF score is interesting in that it implies that subjective similarity between author and followers might cause retweeting, and we call this phenomenon "tight homophily" of following relation because it requires both topic homophily and opinion homophily.

5.3 Performance of Retweeting Prediction

The main purpose of retweeting analysis is to help users find interesting information from the overwhelming information streams. Retweeting is an important signal of interestingness because users are prone to broadcast their favorite messages to their followers. Thus, the performance of retweeting prediction is a suitable evaluation for the utility of subjectivity model in retweeting analysis problem. In our experiment, we evaluate the subjectivity model in supervised machine learning framework.

As Section 4.2 introduced, the retweeting analysis problem could be formulated as a quadruple $< f, a, c, r_{fam} >$. For retweeting prediction, we need to estimate the label r_{fam} when $m, a, and f$ are known. There are 5,214 retweeters in our dataset who retweet at least one target tweet, so we extract 5214 quadruples as positive instances with their label $r_{fam} = 1$. For the other 40,317 non-retweeters, we also extract quadruples as negative instances with label $r_{fam} = 0$. To avoid unbalance bias of training data, we randomly sample 5,214 negative instances into the test dataset.

5.3.1 Comparison With Other User Models

Firstly the comparison between our model with other UGC-based user models (TF-IDF model [40], entity-based model

and hashtag-based model [1]) in retweeting prediction is investigated. As defined in Section 5.2, the six similarities derived from our model are used for comparison, because they model different factors that influence retweeting baehavior. For the comparing models, cosine similarities are calculated between tweets and followers. We use the logistic regression classifier of Scikit-learn machine learning package [51], with 5-fold cross-validation on our balance dataset. Accuracy is our evaluation metric. Performances of our model and all other models are shown in Figure 4. Figure 4 also

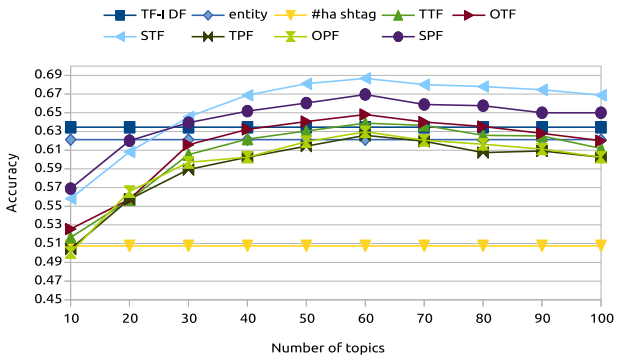

Fig. 4 Comparison of different models.

shows that the impact of topic number of LDA on the predicting accuracy, our model arrives its peak when the number is set to 60, so we fix the topic number as 60 in all our experiment.

As Figure 4 illustrates, the best accuracy of 68.67% is achieved by the STF (Subjective similarity between tweet and followers). The accuracies of TF-IDF model and entity-based model are 63.45% and 62.12%, which are very close to TTF (Topic similarity between Tweet and Followers, 63.88%) and OAF (Opinion similarity between Publisher and Followers, 62.96%). While for hashtag-based model, its accuracy is 50.76% , which is only a little better than random selection (50%) but not significant. The reason might lie in a very low usage of hashtag in our data. The accuracies of the other three model are OTF (Opinion similarity between Tweet and Followers, 64.80%), TAF (Topic similarity between Publisher and Followers, 62.58%) and SAF (Subjective similarity between Publisher and Followers, 66.95%) model. The results show that subjectivity model can better help understanding retweeting behavior than the other user models.

5.3.2 Comparison with Other Factors

In this section, we feed the six similarities of our model as features into a retweeting classification framework to verify the effectiveness of subjectivity model. We compare the performance of our model with method of Luo *et al.* [40] which uses four feature families: Retweet History (follower who retweeted a user before is likely to retweet the user again), Follower Status (for a follower, the number of tweets, followers, friends, being listed and whether he is verified), Follower Active Time (the time users interact with others) and Follower Interests (common interests between tweet and followers, TF-IDF model).

Table 2 Prediction Accuracy of Different Models. Significant improvement over baseline with star($*$) and LUO' model with dagger($\ddagger$) ($p<0.05$).

Feature Set	Accuracy(%)	
RB	60.85	
LUO	68.76 $*$	
SM6	69.12 $*$	
LUO($\ominus$)+TTF	69.20 $*$	
LUO($\ominus$)+TAF	71.04 $*$	$\ddagger$
LUO($\ominus$)+OTF	71.88 $*$	$\ddagger$
LUO($\ominus$)+OAF	70.27 $*$	
LUO($\ominus$)+STF	72.86 $*$	$\ddagger$
LUO($\ominus$)+SAF	72.05 $*$	$\ddagger$
LUO($\ominus$)+All	72.93 $*$	$\ddagger$

We use LinearSVM of Scikit-learn package to build a retweeting prediction framework, leveraging two different features sets. One includes the six features derived from subjectivity model (marked as "SM6"). The other is the feature set from Luo *et al.* [40] (marked as "LUO"). We use the same dataset as Section 5.3.1 with 5-fold cross-validation, and accuracy as evaluation metric. In addition, we set a baseline (marked as "RB"), for which followers who have retweeted the author's previous tweets are predicted as retweeters of current tweet. The result is listed in Table 2. The accuracy of baseline is 60.85%, and two prediction models (LUO and our SM6) both outperform the baseline significantly. But the prediction model based on our feature set shows no significant improvement over LUO feature set. The reason might be that our model only tries to reflect the retweeting motivation of users based on content, whereas other important factors associated with retweeting behavior are not considered, such as network topology and tweeting habit of the user, etc.

Since it is proved that subjectivity model outperforms TF-IDF model in Section 5.3.1, which is used in LUO feature set, we propose that retweeting prediction performance could be improved by using features derived from subjectivity model. As denoted by "LUO($\ominus$)+" in the table, the Follower Interests features of LUO are replaced with our six features one by one. The accuracies are all improved. It shows that our model is of great importance for retweeting prediction. Noticing that, the most significant improvement (LUO($\ominus$)+STF, 72.86% versus 68.76%) is the subjective similarity feature between tweet and followers, which

verifies our assumption that subjective resonance between tweet and followers can be considered as the underlying reason that elicits retweeting behavior. Besides, the improvement by adding subjective similarity features between aauthor and followers (LUO(⊖)+SAF, 72.05% versus 68.76%) is also obvious in that the resonance between author and follower indicates the tight homophily between them. Finally, the last row of table is the complete combination of two sets of features (LUO(⊖)+All) by adding all six features into LUO feature set. The performance shows no significant improvement over adding STF feature only, in that subjectivity model combines both topic and opinion information, and STF is a integral feature to model both topic similarity and opinion similarity between tweet and followers, so it is redundant to add other separate parts.

5.4 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 for one of the 500 target tweets, its author, and two followers (one retweeter, the other non-retweeter) are shown as Figure 5. 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 6 shows top words of the 14^{th} topic, the tweets of author and two followers in a word cloud[2]. Content of the tweet is:

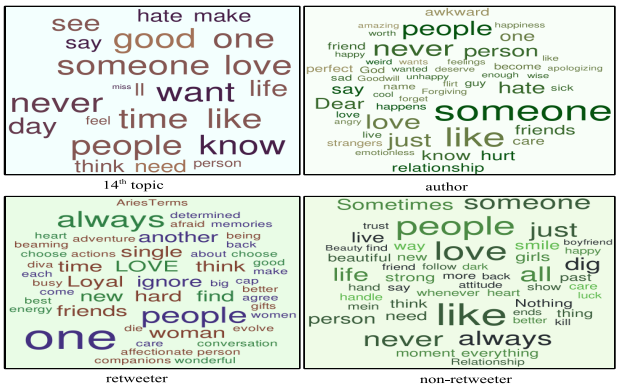

Fig. 6 Word cloud of 14^{th} topic, author and followers.

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

The topic of this tweet is about "love between people" and the opinion is neutral, which is in accordance with the 14^{th} topic word cloud in Figure 6 and subjectivity model of tweet in Firgure 5. The author concentrates on the 14^{th} topic with 208 tweets, and his opinions are mainly neutral (as Figure 5, 6 demonstrate). As for two followers, the retweeter has published 250 tweets about two topics (the 14^{th} and 52^{nd} topic) uniformly and his opinions towards the two topics are mainly neutral. 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 and his opinion is positive. Although two followers have same interest (the 14^{th} topic), 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.

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

6 Conclusion

In this paper, we propose a subjectivity model to analyze user retweeting behavior on Twitter. We assume that retweeting should be elicited by the subjective resonance between the tweet and its followers. We define the subjectivity model formally as the combination of topics and opinions, and we put forward an algorithm to establish the subjectivity model leveraging statistical topic model and sentiment analysis techniques. We demonstrate the effectiveness of our model for retweeting analysis problem and show that subjectivity model is able to reach better understanding of retweeting behavior.

Our future work mainly lie in two directions. Firstly, the subjectivity model is established through simple combination of topics and opinions. 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 subjectivity model to other social media analysis task such as connection prediction and friend recommendation.

References

1. Abel, F., Gao, Q., Houben, G.J., Tao, K.: Analyzing user modeling on twitter for personalized news recommendations. In: User Modeling, Adaption and Personalization, pp. 1–12. Springer (2011)
2. Ahmed, A., Low, Y., Aly, M., Josifovski, V., Smola, A.J.: Scalable distributed inference of dynamic user interests for behavioral targeting. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 114–122. ACM (2011)
3. Armentano, M.G., Godoy, D., Amandi, A.: Recommending information sources to information seekers in twitter. In: International Workshop on Social Web Mining. Citeseer (2011)
4. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 36–44. Association for Computational Linguistics (2010)
5. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. the Journal of machine Learning research **3**, 993–1022 (2003)

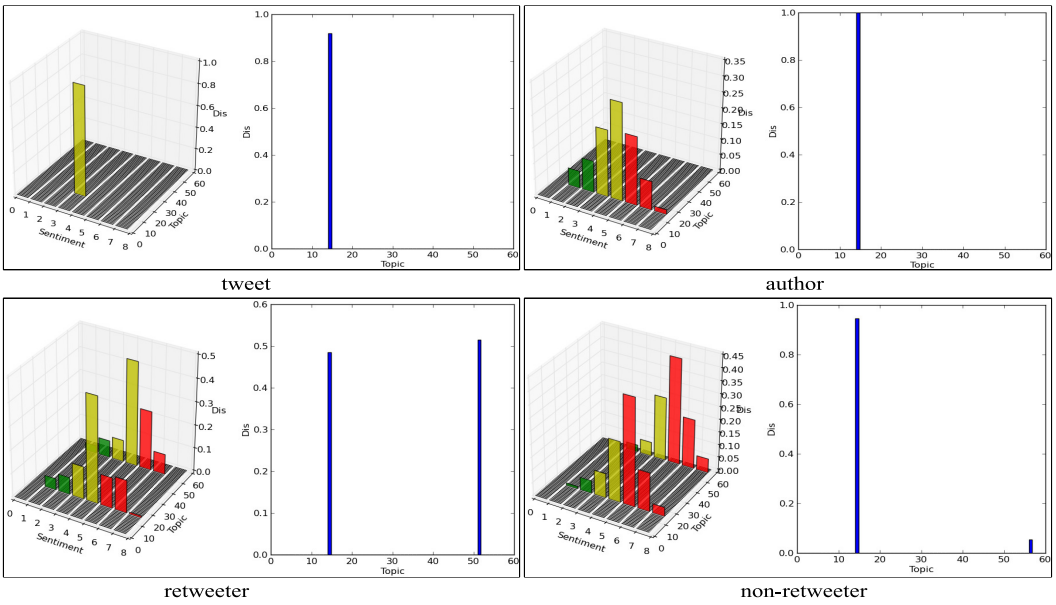

Fig. 5 Subjectivity models of tweet, author and followers.

6. Boyd, D., Golder, S., Lotan, G.: Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In: System Sciences (HICSS), 2010 43rd Hawaii International Conference on, pp. 1–10. IEEE (2010)
7. Brody, S., Diakopoulos, N.: Cooooooooooooooollllllllllll-lll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 562–570. Association for Computational Linguistics (2011)
8. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th international conference on World wide web, pp. 675–684. ACM (2011)
9. Chang, J., Blei, D.M.: Relational topic models for document networks. In: International Conference on Artificial Intelligence and Statistics, pp. 81–88 (2009)
10. Chen, J., Nairn, R., Nelson, L., Bernstein, M., Chi, E.: Short and tweet: experiments on recommending content from information streams. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1185–1194. ACM (2010)
11. Chen, K., Chen, T., Zheng, G., Jin, O., Yao, E., Yu, Y.: Collaborative personalized tweet recommendation. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pp. 661–670. ACM (2012)
12. Chen, Y., Pavlov, D., Canny, J.F.: Large-scale behavioral targeting. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 209–218. ACM (2009)
13. Chua, F.C.T., Lauw, H.W., Lim, E.P.: Generative models for item adoptions using social correlation. Knowledge and Data Engineering, IEEE Transactions on **25**(9), 2036–2048 (2013)
14. Comarela, G., Crovella, M., Almeida, V., Benevenuto, F.: Understanding factors that affect response rates in twitter. In: Proceedings of the 23rd ACM conference on Hypertext and social media, HT '12, pp. 123–132. ACM, New York, NY, USA (2012). DOI 10.1145/2309996.2310017. URL http://doi.acm.org/10.1145/2309996.2310017
15. Cornelius, R.R.: The science of emotion: Research and tradition in the psychology of emotions. Prentice-Hall, Inc (1996)
16. Davidov, D., Tsur, O., Rappoport, A.: Enhanced sentiment learning using twitter hashtags and smileys. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 241–249. Association for Computational Linguistics (2010)
17. Engbert, K., Wohlschläger, A., Thomas, R., Haggard, P.: Agency, subjective time, and other minds. Journal of Experimental Psychology: Human Perception and Performance **33**(6), 1261–1268 (2007). URL http://www.ncbi.nlm.nih.gov/pubmed/18085941
18. Fang, Y., Si, L., Somasundaram, N., Yu, Z.: Mining contrastive opinions on political texts using cross-perspective topic model. In: Proceedings of the fifth ACM international conference on Web search and data mining, pp. 63–72. ACM (2012)
19. Feng, W., Wang, J.: Retweet or not?: personalized tweet re-ranking. In: S. Leonardi, A. Panconesi, P. Ferragina, A. Gionis (eds.) WSDM, pp. 577–586. ACM (2013). URL http://dblp.uni-trier.de/db/conf/wsdm/wsdm2013.html#FengW13
20. Fox, E.: Emotion science cognitive and neuroscientific approaches to understanding human emotions. Palgrave Macmillan (2008)
21. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford pp. 1–12 (2009)
22. Hannon, J., Bennett, M., Smyth, B.: Recommending twitter users to follow using content and collaborative filtering approaches. In: Proceedings of the fourth ACM conference on Recommender systems, pp. 199–206. ACM (2010)
23. Hong, L., Dan, O., Davison, B.D.: Predicting popular messages in Twitter. In: Proceedings of the 20th international conference companion on World wide web, WWW '11, pp. 57–58. ACM, New York, NY, USA (2011). DOI 10.1145/1963192.1963222. URL http://dx.doi.org/10.1145/1963192.1963222
24. Hyman, J.: Three Fallacies about Action. Behavioral and Brain Sciences **23**, 665–666 (2000). DOI 10.1017/S0140525X00223431
25. Ikegami, Y., Kawai, K., Namihira, Y., Tsuruta, S.: Topic and opinion classification based information credibility analysis on twitter. In: Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on, pp. 4676–4681. IEEE (2013)
26. Jenders, M., Kasneci, G., Naumann, F.: Analyzing and predicting viral tweets. In: Proceedings of the 22nd international conference on World Wide Web companion, WWW '13 Companion, pp. 657–664. International World Wide Web Conferences Steering Com-

mittee, Republic and Canton of Geneva, Switzerland (2013). URL http://dl.acm.org/citation.cfm?id=2487788.2488017
27. Jo, Y., Oh, A.H.: Aspect and sentiment unification model for online review analysis. In: Proceedings of the fourth ACM international conference on Web search and data mining, pp. 815–824. ACM (2011)
28. Kang, J.H., Lerman, K., Getoor, L.: La-lda: a limited attention topic model for social recommendation. In: Social Computing, Behavioral-Cultural Modeling and Prediction, pp. 211–220. Springer (2013)
29. Kawamae, N.: Author interest topic model. In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 887–888. ACM (2010)
30. Kawamae, N.: Latent interest-topic model: finding the causal relationships behind dyadic data. In: Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 649–658. ACM (2010)
31. Kumar, R., Tomkins, A.: A characterization of online browsing behavior. In: Proceedings of the 19th international conference on World wide web, pp. 561–570. ACM (2010)
32. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web, pp. 591–600. ACM (2010)
33. Lazarsfeld, P.F., Merton, R.K.: Friendship as a social process: A substantive and methodological analysis. In: M. Berger, T. Abel (eds.) Freedom and control in modern society. Van Nostrand, New York (1954)
34. Lek, H.H., Poo, D.C.: Aspect-based twitter sentiment classification. In: Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference on, pp. 366–373. IEEE (2013)
35. Li, D., Shuai, X., Sun, G., Tang, J., Ding, Y., Luo, Z.: Mining topic-level opinion influence in microblog. In: Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 1562–1566. ACM (2012)
36. Liao, Y., Moshtaghi, M., Han, B., Karunasekera, S., Kotagiri, R., Baldwin, T., Harwood, A., Pattison, P.: Mining micro-blogs: opportunities and challenges. In: Computational Social Networks, pp. 129–159. Springer (2012)
37. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM conference on Information and knowledge management, pp. 375–384. ACM (2009)
38. Liu, Y., Niculescu-Mizil, A., Gryc, W.: Topic-link lda: joint models of topic and author community. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 665–672. ACM (2009)
39. Lu, Y., Zhai, C.: Opinion integration through semi-supervised topic modeling. In: Proceedings of the 17th international conference on World Wide Web, pp. 121–130. ACM (2008)
40. Luo, Z., Osborne, M., Tang, J., Wang, T.: 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, pp. 869–872. ACM, New York, NY, USA (2013). DOI 10.1145/2484028.2484158. URL http://doi.acm.org/10.1145/2484028.2484158
41. Macskassy, S.A., Michelson, M.: Why do people retweet? anti-homophily wins the day! In: L.A. Adamic, R.A. Baeza-Yates, S. Counts (eds.) ICWSM. The AAAI Press (2011). URL http://dblp.uni-trier.de/db/conf/icwsm/icwsm2011.html#MacskassyM11
42. McCallum, A., Corrada-Emmanuel, A., Wang, X.: Topic and role discovery in social networks. Computer Science Department Faculty Publication Series p. 3 (2005)
43. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual review of sociology pp. 415–444 (2001)
44. Mei, Q., Ling, X., Wondra, M., Su, H., Zhai, C.: Topic sentiment mixture: modeling facets and opinions in weblogs. In: Proceedings of the 16th international conference on World Wide Web, pp. 171–180. ACM (2007)
45. Michelson, M., Macskassy, S.A.: Discovering users' topics of interest on twitter: a first look. In: Proceedings of the fourth workshop on Analytics for noisy unstructured text data, pp. 73–80. ACM (2010)
46. Mimno, D., McCallum, A.: Expertise modeling for matching papers with reviewers (2007)
47. Moore, J., Haggard, P.: Awareness of action: Inference and prediction. Consciousness and Cognition **17**(1), 136–144 (2008). DOI 10.1016/j.concog.2006.12.004. URL http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2008-03618-012&site=ehost-live
48. Naveed, N., Gottron, T., Kunegis, J., Alhadi, A.C.: 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). URL http://dl.dropbox.com/u/20411070/Publications/2011-WebSci-Naveed-GKC.pdf
49. Naveed, N., Gottron, T., Kunegis, J., Alhadi, A.C.: Searching microblogs: coping with sparsity and document quality. In: Proceedings of the 20th ACM international conference on Information and knowledge management, CIKM '11, pp. 183–188. ACM, New York, NY, USA (2011). DOI 10.1145/2063576.2063607. URL http://doi.acm.org/10.1145/2063576.2063607
50. Pathak, N., DeLong, C., Banerjee, A., Erickson, K.: Social topic models for community extraction. In: the 2nd SNA-KDD Workshop, vol. 8. Citeseer (2008)
51. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
52. Petrovic, S., Osborne, M., Lavrenko, V.: Rt to win! predicting message propagation in twitter. In: ICWSM (2011)
53. Pfitzner, R., Garas, A., Schweitzer, F.: Emotional divergence influences information spreading in twitter. In: J.G. Breslin, N.B. Ellison, J.G. Shanahan, Z. Tufekci (eds.) ICWSM. The AAAI Press (2012). URL http://dblp.uni-trier.de/db/conf/icwsm/icwsm2012.html#PfitznerGS12
54. Phelan, O., McCarthy, K., Bennett, M., Smyth, B.: Terms of a feather: Content-based news recommendation and discovery using twitter. In: Advances in Information Retrieval, pp. 448–459. Springer (2011)
55. Purushotham, S., Liu, Y., Kuo, C.C.J.: Collaborative topic regression with social matrix factorization for recommendation systems. arXiv preprint arXiv:1206.4684 (2012)
56. Ramage, D., Dumais, S., Liebling, D.: Characterizing microblogs with topic models. In: ICWSM (2010)
57. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: 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, pp. 248–256. Association for Computational Linguistics (2009)
58. Read, J.: Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: Proceedings of the ACL Student Research Workshop, pp. 43–48. Association for Computational Linguistics (2005)
59. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50. ELRA, Valletta, Malta (2010). http://is.muni.cz/publication/884893/en
60. Roberts, K., Roach, M.A., Johnson, J., Guthrie, J., Harabagiu, S.M.: Empatweet: Annotating and detecting emotions on twitter. In: LREC, pp. 3806–3813 (2012)

61. Romero, D.M., Galuba, W., Asur, S., Huberman, B.A.: Influence and passivity in social media. In: Machine learning and knowledge discovery in databases, pp. 18–33. Springer (2011)
62. Saif, H., He, Y., Alani, H.: Alleviating data sparsity for twitter sentiment analysis. CEUR Workshop Proceedings (CEUR-WS. org) (2012)
63. Starbird, K., Palen, L.: (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, pp. 7–16. ACM, New York, NY, USA (2012). DOI 10.1145/2145204.2145212. URL http://doi.acm.org/10.1145/2145204.2145212
64. Stein, D., Wright, S.: Subjectivity and Subjectivisation: Linguistic Perspectives. Cambridge University Press (2005). URL http://books.google.com.hk/books?id=mWlS5Q8uBYcC
65. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.: Probabilistic author-topic models for information discovery. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 306–315. ACM (2004)
66. Stieglitz, S., Dang-Xuan, L.: Political communication and influence through microblogging-an empirical analysis of sentiment in twitter messages and retweet behavior. In: HICSS, pp. 3500–3509. IEEE Computer Society (2012)
67. Stoppard, J.M., Gruchy, C.D.G.: Gender, context, and expression of positive emotion. Personality and Social Psychology Bulletin **19**(2), 143–150 (1993)
68. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: 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), pp. 177–184. IEEE, Minneapolis (2010). DOI citeulike-article-id:7942067. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5590452http://dx.doi.org/10.1109/SocialCom.2010.33
69. Sun, A.R., Cheng, J., Zeng, D.D.: A novel recommendation framework for micro-blogging based on information diffusion. In: Proceedings of the 19th Workshop on Information Technologies and Systems (2009)
70. Sunstein, C.: On Rumors: How Falsehoods Spread, Why We Believe Them, What Can Be Done. Farrar, Straus and Giroux (2009). URL http://books.google.com/books?id=OBoIBIk2qacC
71. Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., Li, P.: User-level sentiment analysis incorporating social networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1397–1405. ACM (2011)
72. Tang, J., Fong, A.: Sentiment diffusion in large scale social networks. In: Consumer Electronics (ICCE), 2013 IEEE International Conference on, pp. 244–245. IEEE (2013)
73. Tchokni, S., Séaghdha, D.O., Quercia, D.: Emoticons and phrases: Status symbols in social media (2014)
74. Thelwall, M.: Homophily in myspace. Journal of the American Society for Information Science and Technology **60**(2), 219–231 (2009)
75. Thelwall, M.: Emotion homophily in social network site messages. First Monday **15**(4) (2010)
76. Wang, C., Blei, D.M.: Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 448–456. ACM (2011)
77. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: B.D.D. 0001, T. Suel, N. Craswell, B.L. 0001 (eds.) WSDM, pp. 261–270. ACM (2010). URL http://dblp.uni-trier.de/db/conf/wsdm/wsdm2010.html#WengLJH10
78. Yan, R., Lapata, M., Li, X.: Tweet recommendation with graph co-ranking. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pp. 516–525. Association for Computational Linguistics (2012)
79. Yang, S.H., Long, B., Smola, A., Sadagopan, N., Zheng, Z., Zha, H.: Like like alike: joint friendship and interest propagation in social networks. In: Proceedings of the 20th international conference on World wide web, pp. 537–546. ACM (2011)
80. Yang, Z., Guo, J., Cai, K., Tang, J., Li, J., 0007, L.Z., Su, Z.: Understanding retweeting behaviors in social networks. In: J. Huang, N. Koudas, G.J.F. Jones, X. Wu, K. Collins-Thompson, A. An (eds.) CIKM, pp. 1633–1636. ACM (2010). URL http://dblp.uni-trier.de/db/conf/cikm/cikm2010.html#YangGCTLZS10
81. Zhang, K., Downey, D., Chen, Z., Xie, Y., Cheng, Y., Agrawal, A., Liao, W.k., Choudhary, A.: A probabilistic graphical model for brand reputation assessment in social networks. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 223–230. ACM (2013)
82. Zhao, T., Li, C., Ding, Q., Li, L.: User-sentiment topic model: refining user's topics with sentiment information. In: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, p. 10. ACM (2012)
83. Zhou, D., Manavoglu, E., Li, J., Giles, C.L., Zha, H.: Probabilistic models for discovering e-communities. In: Proceedings of the 15th international conference on World Wide Web, pp. 173–182. ACM (2006)