Resonation Elicits Diffusion: modelling Subjectivity of Users and Tweets for Retweeting Analysis

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 aspect of a user to retweet a message. Subjective nature of human is the underlying motivation of diverse social behaviors including information diffusion, and subjective resonation triggered by topic and opinion similarity between tweets and users will elicit retweeting behaviors. In this paper, in the light of psychological theory, we put forward that a tweet is more likely to be retweeted by a user because of similar subjectivity, and propose a subjective model to combine both the topics and opinions to model subjectivity of users and tweets. With state-of-the-art topic model and sentiment analysis techniques, we establish subjective 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 influence on retweeting behavior and effectiveness in the retweeting prediction performance. Specifically, we demonstrate that subjective similarity is the most distinguishable factor with largest difference between retweeted and unretweeted users; subjective model outperforms other models(entity-based, hashtag-based models, etc) in predicting rewteeting behavior; and features derived from subjective model gives the most significant improvement over a off-the-shelf predicting model considering network context and meta factors of users and tweets in a rewteeting classification framework.

Categories and Subject Descriptors

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

General Terms

Model, Experimentation

Keywords

Twitter, subjectivity, retweet, LDA, sentiment analysis

1. INTRODUCTION

Twitter is well-known for its freedom of publishing short message (i.e. tweet) within a limited length of 140 characters, and viral spreading of information across complex social networks. Since launched in March 2006, the service rapidly has gained worldwide popularity, with over 500 million registered users at the end of year 2012, who generated over 340 million tweets per day[1]. In addition to large amounts of user-generated content, 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 people's social behaviors on a scale that has never been possible [32].

Information diffusion is a lasting research problem for social scientists who might want to study the problem on Twitter in that, retweeting convention and complex networks of Twitter provide an unprecedented mechanism for the spread of information despite the restricted length of tweets [15]. Actually almost a quarter of the tweets published by users are retweeted from others [39]. Therefore, it is important to understand how retweeting behavior works which can help study information diffusion on Twitter.

Although several studies on retweeting have concentrated on analyzing retweeting habits and influencing factors [4, 17, 33], 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 this process is to estimate whether a tweet contains information interested to the user who might find it worthy to be shared. Therefore modelling the interests of users provides an important perspective for retweet behavior analysis. In this study we focus specifically on how information spreads on Twitter by exploring the retweet behavior from the user modelling perspective.

Previous studies on retweeting behavior have shown that an enriched user model gives coherent and consistent explanation for retweeting motivation[1, 19, 9]. Specifically, researchers have tried to model users from four types of information: profile features ("**Who you are**"), tweeting behavior ("**How you tweet**"), linguistic content ("**What you tweet**") and social network ("**Who you tweet**") [25]. Despite demographic profile, tweeting habits and network structure might determine 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.* [26]

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

and Hong *et al.* [11] found whether a tweet will be propagated largely depends on its identification with the topics of interest 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 freely. Existing researches demonstrated that user-generated content with rich sentimental information can trigger more attention, feedback or participation [32], and tweets with high emotional diversity have a better chance of being retweeted [27]. Until now, most studies have tried to find whether and how sentiment of a tweet will influence its spreading, while none of them realize that although users receive thousands of tweets about different topics every day, whether a tweet will be retweeted will depend on the subjective choice of users.

Philosophically, subjective initiative nature of human determines that his behavior pattern is subjectivity driven. Psychological researchers have identified subjectivity as the underlying factor that influences the decision-making about taking what activities to process incoming stimul [21]. According to theory of Biased Assimilation, people are prone to choose and diffuse information according to their own biased subjectivity [14, 34]. In this study we explore the textual information of Twitter to model the subjectivity of tweets and users, and investigate whether the subjective model could benefit the retweet behavior analysis. Intuitively, subjectivity can be represented as topics and opinions articulated in the information generated by users on Twitter. Technically, We use the state-of-the-art Latent Dirichlet Allocation topic model to find the topics users are talking about, and sentiment analysis techniques to determine user's opinions towards these topics from user-generated content simultaneously. We evaluate subjective model in the retweeting analysis problem to verify its influence on retweeting behavior. Modelling subjectivity on Twitter is a challenging task because of the sparsity of textual information, the dynamics of topics and opinions, and the usages of informal language. 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 subjective 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 subjective 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 subjective model for users and tweets which model both the topics and opinions simultaneously.
- Based on state-of-the-art topic model and sentiment analysis techniques, we establish subjective model from user-generated content on Twitter and combine it with the retweeting behavior analysis problem.
- We systematically evaluate the impact of subjective model on retweeting behavior and demonstrate that it plays important roles in that it outperforms other models in rewteeting prediction and gives the most significant improvement over a off-the-shelf predicting model in a rewteeting classification framework.

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

2. RELATED WORK

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

2.1 Retweet behavior analysis

A large body of studies have analyzed characteristics of retweeting, examining factors that lead to increased retweetability and designing models to estimate the probability of being retweeted.

As for factors influencing retweetability, Suh *et al.* [33] found that tweets with URLs and hashtags were more likely to be retweeted, and there was a strong linear relationship between the number of followers and the likelihood that the tweet be retweeted. Macskassy and Michelson [19] studied a set of Twitter users over a period of a month found that models derived from tweets content could explain most of retweeting behaviors. Comarela *et al.* [7] found previous response to the tweeter, the tweeters' sending rate, the freshness of information, the length of tweet could affect followers' response to retweet. Starbird and Palen [30] addressed specifically the retweeting mechanism during crises and found that tweets with topical keywords were more likely to be retweeted.

There were also many works extending the analysis to build retweeting prediction model. Osborne and Lavrenko [26] 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 the dominance of social features, while tweet features added a substantial boost to the performance. Jenders *et al.* [15] analyzed the "obvious" and "latent" features from structural, content-based, and sentimental aspects of both tweets and users, with respect to their impact on the spread of tweets. They found a combination of features covering all aspects was the key to high prediction quality. Naveed *et al.* [23, 22] introduced interestingness as static quality measure to capture the static content quality of tweets, and quantified it based on such features as emoticons, sentiments and topics a tweet contains, then trained a logistic regression model to predict the probability of retweet for an individual tweet. Feng and Wang [9] built a graph made up of users, publishers and tweets nodes with all sources of information incorporating into nodes and edges, and proposed a feature-aware factorization model to rerank the tweets according to their probability of being retweeted. Pfitzner *et al.* [27] proposed a new measure called emotional divergence to evaluate the retweet probability of a tweet and showed that highly emotional diverse tweets can have up to almost five times higher chances of being retweeted.

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

2.2 User profile and user modelling

With the popularity of social media, researchers have begun to pay close attention to the massive amount of data generated by users, and put forwards several techniques to model users on the data. These studies provide researchers with insights into user online behaviors.

Hannon *et al.* [10] proposed that Twitter users can be modeled by the tweets and the relation of Twitter social network. They found that content-based approach could find similar users who are "distant" without follow relations based on interests extracted from the content of tweets. Macskassy and Michelson [19] discover user's topics of interest by leveraging Wikipedia as external knowledge to determine a common set of high-level categories that covers entities in tweets. Ramage *et al.* [28] made use of topic models to analyze Twitter content at large scale and at the level of individual users with 4S dimensions, showing improved performance on tasks such as post filtering and user recommendation. These efforts about user modelling on Twitter have simply built model for each user by extracting keywords, entities, categories or latent topics from tweet content.

Some researchers argued that user behavior could easily be affected by some external factors other than user interest. Xu *et al.* [38] 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 [25] proposed a most comprehensive method to model Twitter user for user classification. They focused on richer feature sets such as features derived from topic models, tweet sentiment and explicit follower-followed links, etc. Their work confirmed the value of in-depth features by exploiting the user-generated content, which reflect a deeper understanding of the Twitter user and the user network structure.

As introduced in section 1, previous researches have tried to model users from four types of information: profile features, tweeting behavior, linguistic content and social network. Some studies perceived that the implicit features articulated in the user-generated content play an important role in user behavior analysis, and they have proposed diverse techniques to capture such in-depth features to model user's interest. Additionally, a few of work identified the correlation between sentiment of users and their behaviors, but they all failed to model subjectivity of a user as a whole. Motivated by the observation, we firstly put forward subjective model to combine both interests and opinions to model a holistic user.

2.3 Sentiment Analysis

Sentiment analysis is a popular research area for many years. Previous research mainly focused on reviews or news comments. Recently, the research began to pay more and more attention to social media such as Twitter.

Hu *et al.* [12] interpreted emotional signals available in social media data for unsupervised sentiment analysis by providing a unified way to model two main categories of emotional signals: emotion indication and emotion correlation. Jiang *et al.* [16] focused on target-dependent Twitter sentiment classification, they proposed a method to improve target-dependent Twitter sentiment classification by taking target-dependent features and related tweets into consideration. Asiaee T. *et al.* [2] presented a cascaded classifier framework for per-tweet sentiment analysis by extracting tweets about a desired target subject, separating tweets with sentiment, and setting apart positive from negative tweets. Hu emphet al. [13] extracted sentiment relations between tweets based on social theories, and proposed a novel sociological approach to utilize sentiment relations between messages to facilitate sentiment classification and effectively handle noisy Twitter data. Motivated by sociological theories that humans tend to have consistently biased opinions, Calais Guerra *et al.* [5] addressed challenges of topic-based real-time sentiment analysis by proposing a novel transfer learning approach with a suitable source task of opinion holder bias prediction. Thelwall *et al.* [36, 35] designed SentiStrength, an algorithm for extracting sentiment strength from informal English text by exploiting the grammar and spelling styles in typical social media text. In this paper, we adopt SentiStrength for sentiment analysis to build our subjective model, as a finer grain sentiment strength could give us more detailed opinion of users than binary polarized sentiment.

3. SUBJECTIVE MODEL

In this section, we firstly give the definition of subjective model. Then we describe the method of building subjective model. Finally, we combine subjective model with the retweeting analysis problem to find the in-depth reason of retweeting behavior.

3.1 Definition

Subjectivity has been extensively studied by psychologists to characterize the personality of a person based on his historic behaviors and remarks [8]. Linguists define the subjectivity of language as the speakers always show their perspectives, attitudes and sentiments in their discourses [31]. And as a free platform, social media provides users a place to express their opinions towards topics of interest to show their personal subjectivity by publishing short messages. Therefore, for the term "subjectivity" on social media, we refer to both topics of interest and opinions towards these topics articulated in the user-generated content, so we model subjectivity not only by topics users care about, but also by "**what they think about the topics**". The user-generated content of social media have provided massive language resources to find both the topics and opinions needed to model subjectivity. Here we firstly give our definition of subjective 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 O for evaluating their opinions towards these topics. As for O, 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 (SUBJECTIVE MODEL FOR USER). *The subjective model of a user* $\mathrm{u} \in \mathrm{U}$ *is the combination of a set of topics* $\{\mathrm{t_i}\,(\mathrm{i} \in \{1 \cdots \mathrm{n}\})\}$ *the user talks about in a topic space* T *and the user's opinion* $\mathrm{o_i}$ *towards each topic* $\mathrm{t_i}$.

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

where:

- *with respect to the given user* u*, for each topic* $\mathrm{t_i} \in \mathrm{T}$*, its weight* $\mathrm{w_u}\left(\mathrm{t_i}\right)$ *represents the distribution of the user's interests on it.*
- *opinion* $\mathrm{o_i}$ *of user towards topic* $\mathrm{t_i}$ *is a target-dependent sentiment distribution* $\mathrm{d_{u,t_i}}\left(\mathrm{o_i}\right)$ *over sentiment valence space* O.

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

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

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

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

3.2 Retweeting Analysis Problem Statement

Retweeting is the core mechanism of information diffusion on Twitter. It has been demonstrated that any retweeted tweet is able to reach an average of 1,000 users no matter how many followers the publisher has [17]. Retweeting function gives every user the power of spreading information broadly, and every user has the power to dictate which information is important and should spread by pushing retweeting button. Many factors have been proved to influence retweeting behavior [33, 19, 7], however few researches have investigated the subjective motivation of a user to retweet a message. Therefore we will study whether subjective model can help understand underlying reasons of a user's retweeting behavior in the paper.

In fact the likelihood of a tweet to be retweeted depends on both context constraints and its content. Context such as the author's position in the network or the time of day a tweet is published must influence whether the tweet will be retweeted, in that a tweet with only few or passive followers is less likely to be retweeted, and tweets published in the night have less chance to be retweeted than daytime. Apart from context constraints, a tweet is more likely to be retweeted by subjective users who find it worth to. Therefore, we are not interested in modelling the tweet by itself just as other researches [23, 22, 27], but how the tweet content resonate with the individuals who might want to pass it on. 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 user-generated content.

Actually, neither contextual factor 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 retweeting a tweet lies in that the user considers only the tweet content arousing resonation with him without considering context settings.

In the context of Twitter, the "following" relationship is a strong indicator of a phenomenon called "homophily", which has been observed in many social networks. Homophily is a phenomenon that people connected in a social network "are homogeneous with regard to many socio-demographic, behavioral, and intra-personal characteristics" [20]. In other words, homophily implies that a user follows another user because he is interested in what another user tweets about, and another user follows back because he finds they share similar interests. According to the principle of homophily, we put forwards the concept of **Local Topic Space**, which could be defined as:

DEFINITION 3 (LOCAL TOPIC SPACE). *In a local network consisted of a user and followers, all users concentrate on limited topics derived from the content generated by them, 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, P, C denote the follower set, publisher set and tweet set respectively. For each tweet c ($c \in C$) and its listener f ($f \in F$), we can define a quadruple $< f, p, c, r_{fpc} >$ where:

- p ($p \in P$) is the publisher of the tweet c and f ($f \in F$) is a follower of publisher p.
- r_{fpc} is a binary label indicating whether c is retweeted by f.
- Our work focuses on using subjective model to analyze the relation between the subjectivity of a user and his retweeting behavior. Hence we transform the quadruple into the Local Topic Space T, which is derived from the content generated by publisher p and followers F, and represent f, p, c with their subjective models to analyze their relations with the label r_{fpc}.

3.3 Establishment of Subjective Model

According to definition of subjective model, there are two distributions to model the subjectivity: one is topic distribution and the other is opinion distribution for each topic. Both of them need to be inferred from historic content produced by users. However, content analysis on Twitter has some challenges: 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 models 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 rule-based sentiment analysis techniques, we establish subjective model by finding topics and determining opinions in unsupervised way simultaneously.

3.3.1 Topic Analysis for Tweets

Connection relation between users indicates their common interests in a Local Topic Space. However, the topics of a tweet are latent features and have to be inferred by analyzing its content. Previous studies have tried to infer topics by finding key words [6], extracting entities [1] or linking tweets to external knowledge categories [19], however, the sparsity is a main problem for these methods to model the users' interests because even users have common local topics they still might refer to a topic with different vocabulary. Recent works show that topic models such as **Latent Dirichlet Allocation (LDA)** model and its extensions[3, 37] have been efficient ways to characterize latent topics of large volum 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 a publisher and followers 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 that 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 subjective model.

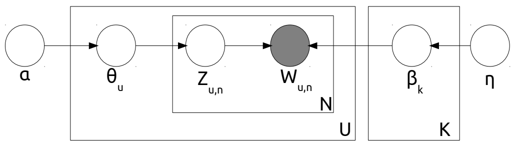

Figure 1: Plate illustration of the user-level LDA model.

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 generative process works as follows:

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

The parameters θ_u and each β_k can be estimated by Gibbs sampling or variational inference. We use variational inference-based topic model package Gensim [29].

3.3.2 Sentiment Analysis for Tweets

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 use machine learning or rule-based approaches. Machine learning approaches often needs 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 [36, 12].

The rule-based SentiStrength package has been built especially to cope with sentiment analysis in short informal text of social media [36]. 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 and a negative 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 of a user's subjective model. For the convenience of distribution calculation, we map the output of SentiStrength to single-scaled sentiment valence space $[0, 8]$ as follows:

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

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

3.3.3 Concrete Subjective Model

With statistical topic analysis and sentiment analysis described above, we can concrete subjective 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 $C_u = \{c_i | i \in [1, \cdots, N]\}$. Each C_u is concatenated to a single document d_u to construct Local Topic Space $T = \{t_i | i = 1, \cdots, K\}$. In the Local Topic Space T, a topic model is built with parameter θ representing the distribution of each user over topics he tweets about, and parameter β represents the distribution of each topic over the vocabulary of all tweets. SentiStrength is applied to each tweet c in collection C_u and outputs sentiment strength s_c for tweet c. We build the subjective model of user u as follows:

- Firstly, for user u, the corresponding component θ_u is the his topic distribution in the Local Topic Space T. $p(z_u|\theta_u)$ could be regarded as the weight of subjective model $w_u(t_i)$, and topics he tweets about are $Z_u = \{z_u | p(z_u|\theta_u) > 0\}$.
- Secondly, the topic model of Local Topic Space T is applied to each tweet c to find topics of c, which are $Z_c = \{z_c | p(z_c|\theta, \beta, Z_u) > 0\}$.
- Thirdly, the opinion distribution of user u towards topic $t \in Z_u$ could be calculated as:

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

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

$$N_o = \sum_{c \in Cu} I(s_c), \text{ if } s_c = o \& t \in Z_c \tag{5}$$

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

where s_c denotes the sentiment strength of tweet c, and for simplicity, we assume the sentiment of tweet c is related to every topic it talks about in Z_c.

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

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

and accordingly subjective model for tweet c as:

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

3.4 Retweeting Analysis With Subjective Model

We are interested the relationship among subjective model of publisher, follower and tweet. For a tweet c, the corresponding publisher p, and a list of followers $F = \{f_i | i = 1, \cdots, N\}$, for each $f_i \in F$, a tuple $\langle f_i, p, c, r_{fpc}\rangle$ could be defined as Section 3.2. We firstly build subjective model $P(u)$ for each user $u \in F \bigcup p$ and $P(c)$ for tweet c in the Local Topic Space T (defined by users in $F \bigcup p$). We assume that, if a user retweet a message, the user not only finds the topics of tweet is interesting but also share similar opinions towards these topics. We called this process "resonate". With the subjective models built for users and tweets, we could define a similarity measurement to quantify the resonation among them:

$$Sim(c, f_i) = similar(P(c), P(f_i)) \tag{9}$$

according to Equation 7,8:

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

where

- λ is coefficient used to control the proportions of topic similarity and opinion similarity in the holistic subjective similarity. We initiate it by setting $\lambda = 0.5$ (You'd better do some experiment how to set λ).
- $Dist$ is the similarity measurement between two distribution, we use *Cosine Similarity* in our research (Why did you use *Cosine Similarity* ? How about KL divergence).

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

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

4. EXPERIMENT

In this section, we investigate whether subjective model can help retweet analysis.

Table 1: Retweet Dataset Statistics

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

4.1 Dataset You should give more detail about the data, since nobody like to read original paper!!!

We adopt an off-the-shelf Twitter dataset[18]. For the dataset, 500 randomly selected English tweets which had been retweeted at least once were used as test tweets, then each test tweet was chosen as starting point to collect data of its publisher and followers. Summary statistics of the dataset are listed in Table 1.

4.2 Example of Subjective Model The position of this subsection is strange, you can put it into introduction, the beginning of section 3 or last

As the core of our work, how to build subjective model has been elaborated in section 3.3.3, in this section we give an qualitative description about subjective model and its ability in explaining the retweet behavior with an intuitive example.meaningless There are 500 test tweets, 500 corresponding publishers, 4,5531 followers and 6,277,736 published tweets in the dataset You should introduce it in dataset section , the relations are illustrated in Figure 2.

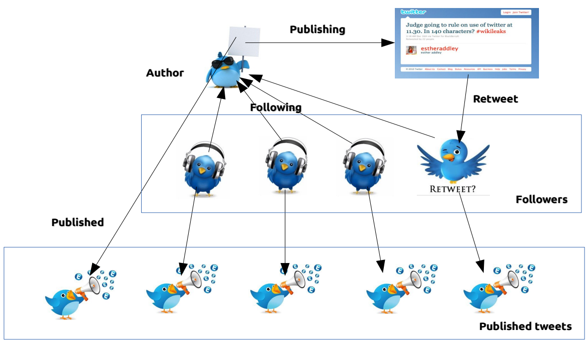

Figure 2: Illustration of dataset structure.

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

Figure 3 plot top words of the 14th topic.

Figure 4 shows the tweets of publishers and two followers in a word cloud[2].

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

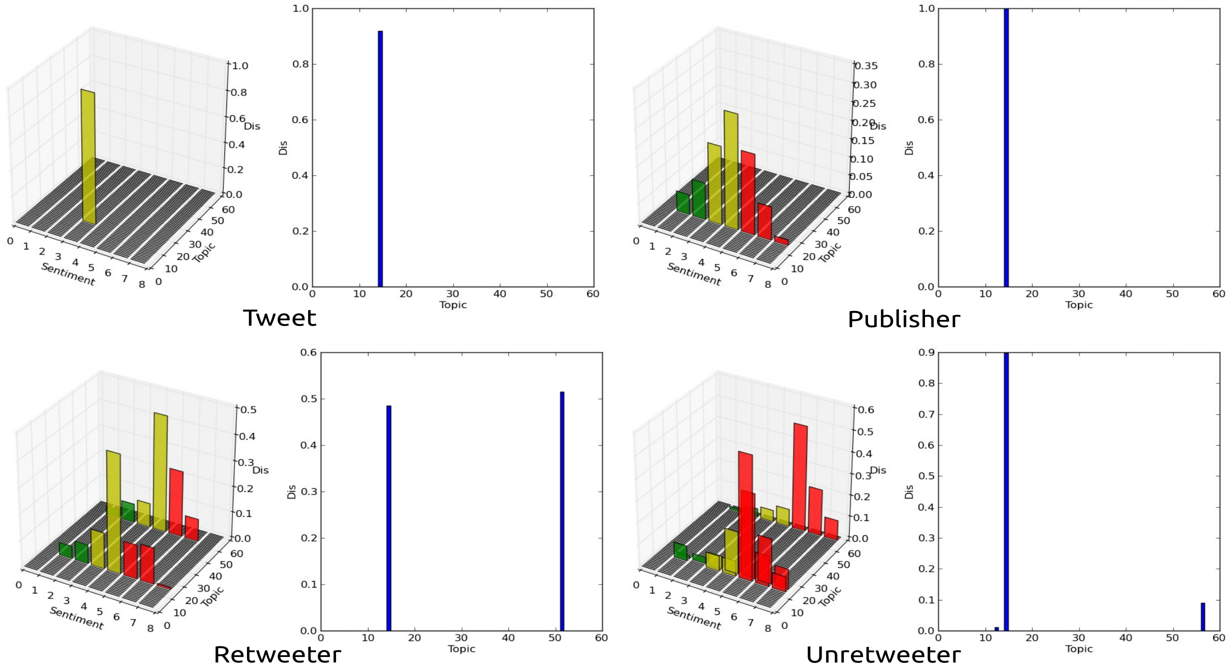

Figure 3: Subjective model examples.

Figure 4: Word cloud of 14th topic, publisher and followers

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

The topic of this tweet is about "love between people" and the opinion is neutral, which is in accordance with the 14th topic word cloud in Figure 4 and subjective model of tweet in Firgure 3. The publisher of tweets also talks about "love between people", and his opinions are mainly neutral (see Figure 3). As for two followers, the "retweeter", who retweet the example tweet, has published tweets about two topics (the 14th and 52nd topic) uniformly and his opinions towards the two topics are mainly neutral. While the other one, who did not retweet the example tweet (we call "unretweeter"), has also talked about two topics (14th and 56th topic), but he is mainly interested in "love between people" topic and has positive opinions. Although two followers have same interest (the 14th topic), the difference of their opinions towards the topic elicits their different retweet behavior, which verifies our subjective model can help understand the retweet behavior.

word cloud.

4.3 Influence of Subjective Model on Retweet

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

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

To analyze the influence of different factors on retweeting, we averaged six similarity scores on 5214 followers who retweet the testing tweets and 5214 randomly selected followers who do not retweet separately. Figure 5 shows the comparing result.

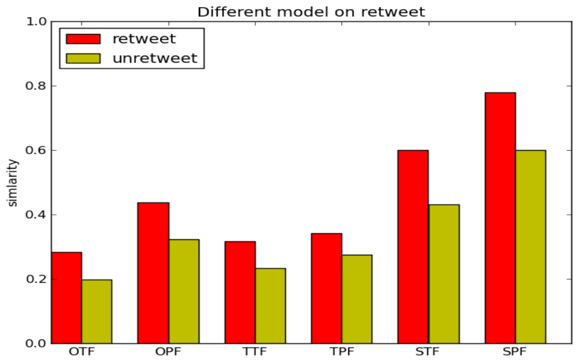

Figure 5: Influence of different factors on retweet.

As the figure demonstrated, on average, the similarities scores of retweeted followers are clearly higher than unretweeted followers for all six factors. It show that our assumption is reasonableGive the conclusion directly not "our assumption". Specifically not listing,

- 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[19, 9];
- 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[27, 22] have shown that sentiment in tweet have impact on retweet, most of them don't consider the opinions of followers and the opinion similarity betweet tweet and its followers;
- STF score shows the subjective model we put forward is the most distinguishable feature among the six factors with the largest difference between retweeted and unretweeted followers, which proves the importance of subjectivity;
- TPF score gives another perspective for retweet from the topic similarity between tweet publisher and followers, indicating that followers are more likely to retweet those whose interests are similar, which verifies the homophily principle of retweet relation;
- OPF score indicates that similar opinions for common topics of interest also influence followers' decision of retweeting anothor user, which may be another proof of homophily of retweet relation.
- SPF score is interesting in that it implies that subjective similarity between user and follower might cause retweet, and we call this phenomenon "tight homophily" because it requires both topic homophily and opinion homophily.

The six scores used to model the factors that influence retweet 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 TPF, OPF and SPF, which is indirect and implicit by modelling the tweet publisher and followers. The two aspects reflect properly the information sharing and diffusion structure of Twitter at micro-level as illustrated in Figure 2. (I do not understand.)

4.4 Performance of Retweet Prediction

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

As Section 3.2 introduced, the retweet problem could be formulated as a tuple $< f, p, c, r_{fpc} >$. In the prediction experiment we need to estimate the label r_{fpc} when c, p, and f are known. There are 5,214 users in our dataset who retweet testing tweets, so we extract 5214 tuples as positive instances with their label $r_{fpc} = 1$. The other 40,317 users who do not retweet any testing tweets are also extracted to form negative tuples with label $r_{fpc} = 0$. Avoiding unbalance bias of training data, we randomly sample 5,214 negative instances into the final dataset.

4.4.1 Comparison With Other User Models

Firstly the comparison between our model with other user models (TF-IDF model [18], entity-based model and hashtag-based model [1]) in predicting retweet are investigated. As for our model, the six parts defined above are used for teasing, because they model different factors that influence retweet. For the comparing models, cosine similarities are calculated between tweets and their followers. We use the logistic regression classifier of Scikit-learn machine learning package [24] for training, with 5-fold cross-validation on our balance dataset. Accuracy is our evaluation metric.

Figure 6 gives the performances of our model and all other models.

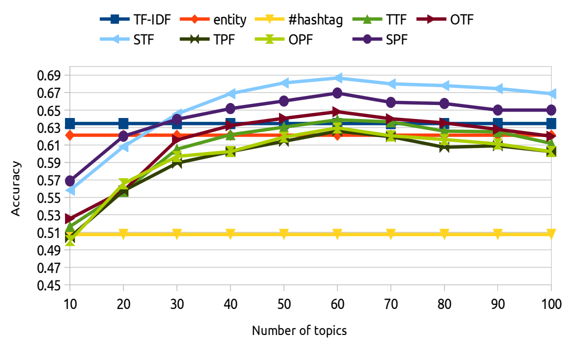

Figure 6: Comparison of different models.

The highest accuracy of 68.67% is the STF (Subjective similarity between Tweet and Followers) model achieved. The accuracies of

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

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

TF-IDF model and entity-based model are 63.45% and 62.12%, which are very close to our TTF (Topic similarity between Tweet and Followers, 63.88%) model and OPF (Opinion similarity between Publisher and Followers, 62.96%) model. While for hashtag-based model, its accuracy is 50.76% , which is only little better than random selection (50%)but not significant. The reason might be that the number of hashtag in our data is not very large. The accuracies of the other three model based on our method are OTF (Opinion similarity between Tweet and Followers) model 64.80%, TPF (Topic similarity between Publisher and Followers) model 62.58% and SPF (Subjective similarity between Publisher and Followers) model 66.95%. The results show that subjective model can better help understand retweet behavior than the other models.

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

4.4.2 *Retweet Classification Evaluation*

In this section, we feed the six parts of our model as features into a retweeting classification framework to verify the effectiveness of our subjective model. We compare the performance of our model with a prediction model of Luo et al. [18] which uses four feature families: Retweet History, Follower Status, Follower Active Time and Follower Interests. You would better introduce these features.

We use LinearSVM of Scikit-learn package to build a retweet prediction model, leveraging two different features sets. One includes the six features derived from subjective model (marked as “SM6”). The other is Luo et al. [18](marked as “LUO”) feature set in which they use “bag-of-words” to model the followers interest. We use the same dataset as introduced in Section 4.4.1 with 5-fold cross-validation, and accuracy as evolution metric. In addition, we set a baseline (marked as “RB”), in which followers who have retweeted the publisher’s previous tweets before are predicted as retweeters for current tweet.

The result is listed in Table 2. The accuracy of baseline is 60.85% Both prediction model based on sets of features (LUO and our SM6) outperform the baseline significantly. But the prediction model based on our feature set shows no significant improvement over the model based on LUO feature set. The reason might be that our model only tries to reflect the retweet motivation of users based on content, whereas other important factors associated with retweet are not considered, such as network context and reading habit of the user. As denoted by “LUO($\ominus$)+” in the table, we combine the two sets of features by replacing the Follower Interests features of LUO model with our six features one by one. The accuracies are all improved. It shoes that our model is of great importance for retweet prediction models. Notice that, the most significant improvement (LUO($\ominus$)+STF, 72.86% versus 68.76%) is the subjective similarity features between tweet and followers, which verifies our assumption that resonation between tweet and the followers elicits retweet behavior(wrong sentence). Besides, the improvement by adding subjective similarity features between publisher and followers (LUO($\ominus$)+SPF, 72.05% versus 68.76%) is also obvious in that the resonation between publisher and follower indicates the tight homophily between them. Finally, the last row of table is the complete combination of two sets of features (LUO($\ominus$)+All) by adding all six features into LUO feature set. The performance shows no significant improvement over adding STF feature only (reason?).

5. CONCLUSION

In this paper, we propose subjective model to analyze user retweeting behavior on Twitter. We assume that retweeting behavior should be elicited by the subjective resonation between the tweet and its followers. We define subjective model formally as combination of topic distribution and opinion distribution, and we concrete subjective model leveraging statistical topic model and sentiment analysis techniques. We demonstrate the effectiveness of our proposed model with retweeting analysis problem and show that this model is able to reach more comprehensive understanding of retweeting.

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

6. REFERENCES

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

[2] A. Asiaee T., M. Tepper, A. Banerjee, and G. Sapiro. If you are happy and you know it... tweet. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 1602–1606, New York, NY, USA, 2012. ACM.

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

[4] D. Boyd, S. Golder, and G. Lotan. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *2010 43rd Hawaii International Conference on System Sciences*, volume 0, pages 1–10, Kauai, HI, 2010.

[5] P. H. Calais Guerra, A. Veloso, W. Meira, Jr., and V. Almeida. From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 150–158, New York, NY, USA, 2011. ACM.

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

'10, pages 1185–1194, New York, NY, USA, 2010. ACM.

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

[8] K. Engbert, A. Wohlschläger, R. Thomas, and P. Haggard. Agency, subjective time, and other minds. *Journal of Experimental Psychology: Human Perception and Performance*, 33(6):1261–1268, 2007.

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

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

[11] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in Twitter. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 57–58, New York, NY, USA, 2011. ACM.

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

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

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

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

[16] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 151–160, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

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

[18] Z. Luo, M. Osborne, J. Tang, and T. Wang. Who will retweet me?: finding retweeters in twitter. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '13, pages 869–872, New York, NY, USA, 2013. ACM.

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

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

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

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

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

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

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

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

[27] R. Pfitzner, A. Garas, and F. Schweitzer. Emotional divergence influences information spreading in twitter. In J. G. Breslin, N. B. Ellison, J. G. Shanahan, and Z. Tufekci, editors, *ICWSM*. The AAAI Press, 2012.

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

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

[30] K. Starbird and L. Palen. (how) will the revolution be retweeted?: information diffusion and the 2011 egyptian uprising. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, pages 7–16, New York, NY, USA, 2012. ACM.

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

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

[33] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *Proceedings of the IEEE Second International Conference on Social Computing (SocialCom)*, pages 177–184, Minneapolis, Aug. 2010. IEEE.

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

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

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

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

[38] Z. Xu, Y. Zhang, Y. Wu, and Q. Yang. Modeling user posting behavior on social media. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 545–554, New York, NY, USA, 2012. ACM.

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