A Systematic Framework for Sentiment Identification by Modeling User Social Effects

Kunpeng Zhang Department of Information and Decision Sciences, University of Illinois at Chicago, USA 60607, Email: kzhang6@uic.edu
Yi Yang Northwestern University, Evanston, USA 60208, Email: yiyang@u.northwestern.edu
Aaron Sun Samsung Research America, San Jose, USA 95134, Email: a.sun@samsung.com
Hengchang Liu School of Computer Science and Technology, University of Science and Technology of China, Suzhou, China 215123, Email: hcliu@ustc.edu.cn

***Abstract*—Social media is becoming a major and popular technological platform that allows users to express personal opinions toward the subjects with shared interests. Identifying the sentiments of these social media data can help users make informed decisions. Existing research mainly focus on developing algorithms by mining textual information in social media. However, none of them collectively consider the relationships among heterogeneous social entities. Since users interact with social brands in social platforms, their opinions on specific topics are inevitably dependent on many social effects such as user preference on topics, peer influence, user profile information, etc. In this paper, we present a systematic framework to identify sentiments by incorporating user social effects besides textual information. We apply distributed item-based collaborative filtering technique to estimate user preference. Our experiments, conducted on large datasets from current major social platforms, such as Facebook, Twitter, Amazon.com, and Flyertalk.com, demonstrate that incorporating those user social effects can significantly improve sentiment identification accuracy.**

***Keywords*—Sentiment, social effects, collaborative filtering, peer influence.**

I. Introduction

Social media has become one of the most popular communication platforms that allow users to discuss and share topics of interest without necessarily having the same geo-location and time. Information can be generated and managed through computers or mobile devices by one person and consumed by many others. Different people can express different opinions on the same topic. People can also express their opinions on multiple topics of interest. A wide variety of topics, ranging from current events and political debate, to sports and entertainment, are being actively discussed on these social forums. For example, Facebook users could comment on or "like" campaigns posted by a company. Twitter users could send tweets with a maximum length of 140 characters to instantly share and deliver their opinions on politics, movies, sports, etc. Some e-commerce platforms, such as Amazon.com, etc. allow users to leave their reviews on products. Some online forums such as Flyertalk.com allow registered users to comment airlines in terms of services, prices, delays, dining, security, etc. The power of social media as a marketing tool has been recognized, and is being actively used by governments, major organizations, schools and other groups to effectively and quickly communicate with a large number of people. Another important metric for business to measure their online reputation is word of mouth publicity. Word of mouth is the process of spreading information from person to person, and is often done through social media networks. Social media is also a good platform to help companies or organizations target potential customers to publish their advertisements based on user historical behavior information.

Identifying sentiments in social media platforms has become an important issue and has attracted a lot of attention. The results of sentiment analysis can help users or managers make informed decisions. For example, marketing leaders or product managers might collect and analyze network information, feedbacks, and comments on their campaigns aiming to adopt efficient advertising strategy and improve product quality. Recently, there has been a number of studies attempting to identify sentiments of social media data [11]–[16]. However, most of these works focus on designing algorithms by mining only textual information in social media. Few works have considered user social effects from heterogeneous social entities, including user preferences on topics, peer influence, user profile information, etc. Sentiment of user-generated content can be affected by these social effects. This work studies to combine them for improving sentiment identification accuracy from large amounts of data.

We believe that incorporating user social effects in sentiment analysis can help reduce biases from individual views. The reason is that users' opinions are usually affected by their social environment. Specifically, 1) Users usually have different preferences on various topics. Sentiments of comments made on different topics should not be equally considered; 2) Based on the herd behavior in psychology, people usually follow what previous users said. Therefore, influence of comments from previous users (peer influence) should also be considered in sentiment identification; 3) From the data, we also found that males express sentiment differently from females.

In this paper, in order to address these problems we propose a model to incorporate each social effect and show the accuracy of sentiment prediction is improved by coupling with previously developed textual analysis algorithms. We conduct our experiments on data from several major types of social platforms, including social networking sites: Facebook, Twitter; online e-commerce sites: Amazon.com; and online forums: Flyertalk.com. Due to the volume of the data, we employ a distributed item-based collaborative filtering technique to calculate user preferences on different topics. Different similarity measurements are tested to obtain a minimum estimation error. To reduce the data sparsity, we use high-level information such as categories on Facebook/Twitter instead of individual pages to represent "items". In summary, the contributions of this paper are as follows.

- We present a distributed item-based collaborative filtering technique to approximate user preferences based on large amounts of user historical activities.
- We develop a model to capture the influence of comments from previous users (peer influence) based on herd behavior in psychology.
- We propose a systematic framework to improve sentiment identification by incorporating user social effects, such as user preference, peer influence, and user profile, in addition to textual sentiments.
- Experiments on four major social media data show that our method can significantly improve sentiment classification accuracy.

The rest of this paper is organized as follows. We list some related work in Section II. We describe the problem statement in Section III. In Section IV, we first present the overall system framework. Then, we show how to estimate user preferences using distributed collaborative filtering techniques, present how to model the peer influence, and describe a textual analysis algorithm. Section V shows experimental data and results, which is followed by the conclusion and future work in Section VI.

II. Related Work

The main research efforts on sentiment include various sentiment identification applications in different domains and sentiment classification [11]–[16]. Most of them deal with one of two problems. The first problem is that researchers took state-of-the-art sentiment identification algorithms to solve problems in real applications such as summarizing customer reviews [2] and finding product features that imply opinions [18]. In [3], authors analyzed sentiments about movies in twitter and attempted to predict box office revenue. Joshi *et al.* used a similar technique to predict box-office revenue of movies using review text [4].

The other problem is to design new sentiment detection, identification, and classification algorithms. These efforts fall into three major categories. 1) Bag-of-Words approaches produce domain-specific lexicons. There is a vast body of research which attempts to incorporate them as features in machine learning models [5], [7], [21]. For example, [26] incorporated Internet languages, such as emoticons to improve the sentiment classification of social media data. 2) Rule-based approaches have also been studied by many researchers. The authors in [22] proposed compositional semantics, which is based on the assumption that the meaning of a compound expression is a function of the meaning of its parts and of the syntactic rules by which they are combined. They have developed a set of compositional rules to assign sentiments to individual clauses, expressions and sentences. 3) Recently, there has been a wide range of machine learning techniques, which classify the whole opinion document (e.g., a product review) as positive or negative [7], [8], [10], [29]. In [22], the authors viewed such sub-sentential interactions in light of compositional semantics, and presented a novel learning-based approach that incorporates structural inference motivated by compositional semantics into the learning procedure. In [7], authors employed machine learning techniques to classify documents by overall sentiments and conducted their experiments on movie reviews and the results show that three machine learning methods they employed (Naïve Bayes, maximum entropy classification, and support vector machines) do not perform as well on sentiment classification as on traditional topic-based categorization. In [6], authors presented a linguistic analysis of conditional sentences, and built some supervised learning models to determine if sentiments expressed on different topics in a conditional sentence are positive, negative or neutral. Several researchers have also studied feature/topic-based sentiment analysis [17]–[21]. Their objective is to extract topics or product features in sentences and determine the associated sentiments. In [5], authors used feature-based opinion mining model to identify noun product features that imply opinions. In [9], authors proposed an approach to extract adverb-adjective-noun phrases based on clause structure obtained by parsing sentences into a hierarchical representation. They also proposed a robust general solution for modeling the contribution of adverbials and negation to the score for degree of sentiment. In [23], authors showed that information about social relationships can be used to improve user-level sentiment analysis. In [30], [31], authors considered network context to model the effect of emotions in sentiment. Our work is different from this in that we capture relationships among social entities and incorporate user social effects into sentiment identification.

Our work is also related to collaborative filtering (CF). CF is a technique by using the known preferences of a group of users to make recommendations or predictions of the unknown preferences for other users. GroupLens [27] first introduced an automated collaborative filtering system using a neighborhood-based algorithm. They used Pearson correlations to weigh user similarity, used all available correlated neighbors, and calculated a final prediction by performing a weighted average of deviations from the neighbor's mean.

III. Problem Statement

Current major social platforms such as Facebook, Twitter, Amazon.com, and Flyertalk.com have common structures which allow users to interact with each other. One user (can also be an administrator for a public page) launches discussions (D^1, D^2, $\cdots$, D^i, $\cdots$, D^d) on various social topics, products, or brands (B^1, B^2, $\cdots$, B^i, $\cdots$, B^b). Each B^i has a specific category C^i. Users with different backgrounds and interests can make their comments, which can be either positive, negative, or objective. They can read comments from previous users. Therefore, they can be influenced by these previous comments, which is called herd behaviors in psychology [24]. Usually, these comments are displayed in multiple pages because they can not be shown in one page due to the size. In addition, users can also use other ways quantitively/qualitatively rate objects (e.g., a 5-star rating on Amazon.com and the "like" button on Facebook). They can help us estimate user preference which consequently affects sentiments. In this paper, the goal is to incorporate these social effects to improve sentiment identification. Slightly more formally, we define our problem as follows.

Input: 1) A discussion thread D^i under a social brand B^i with a topic (usually category) C^i, and the sentiments of all previous textual comments (T_1^i, T_2^i, $\cdots$, T_{m-1}^i) made by users (U_1^i, U_2^i, $\cdots$, U_k^i), where $k \leq m-1$, because each user can make more than one comment.

2) $P_{U_m^i C^i}$: the preference on the current topic C^i from the U_m^i user. All these kinds of preferences are calculated using collaborative filtering described in section IV-A.

Output: The identified sentiment of the m^{th} comment (T_m^i) from some user (U_j^i). This is an iterative process since the first comment.

IV. Methodology

In this section, we first present our systematic overall framework of sentiment identification by incorporating user social effects. Within this system, a support vector machine (SVM) learning model is employed to classify sentiments. Four major features include user preference ("UserPref"), peer influence ("PeerInf"), textual analysis ("TextSent"), and user profile information ("GenCat"). In the following sections, we introduce each in details. First, a distributed item-based collaborative filtering technique is used to approximate user preferences on topics based on large amounts of user historical activities. Then, we analyze how previous comments affect user sentiments and model peer influence (named as "PeerInf"). Then, we introduce our textual analysis algorithm and discuss how user profile information (e.g., gender) affects sentiment classification. Figure 1 shows the framework of our sentiment identification system. The training and testing datasets are randomly generated from our large labeled data pool.

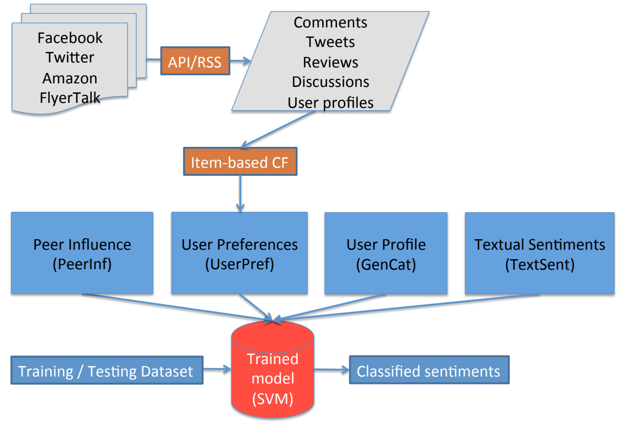


Fig. 1: The general framework of the social sentiment identification system. "TextSent", "UserPref", "PeerInf", and "GenCat" are four major features.

A. *UserPref: User Preference*

The motivation of approximating user preferences ("UserPref") on topics (e.g., sports, TV show, politics, clothing, food/beverages, etc.) is that user preferences can somehow reflect user sentiments. The reason that the user preference approximation is based on user historical activities (likes or ratings instead of comments) is that these quantitative measures can more directly reflect the strength of user preferences. We use collaborative filtering (CF) techniques to approximate user preferences based on the user-item matrix ("item" can be page category on Facebook/Twitter or product category on Amazon.com or airlines on Flyertalk.com).

There are three types of collaborative filtering: user-based, item-based, and content-based. In the user-based collaborative filtering system, especially neighborhood-based, the user-user similarity computation step turns out to be the performance bottleneck, which becomes worse in our case because of a large number of users. One way of ensuring high scalability is to use an item-based approach. One possible way of computing the item similarities is to compute all-to-all similarity and then performing a quick table look-up to retrieve the required similarity values. This method requires an $O(k^2)$ space for k items. Using single CPU to solve this problem would be extremely time consuming. In this paper, we employ a scalable item-based collaborative filtering library implemented by Mahout[1].

The most challenging issue in our item-based collaborative filtering is data sparsity which can be alleviated by reducing the number of columns. Our strategy is to integrate multiple low-level items into less high-level items. For example, "Mac", "iPhone" have the same category on Facebook: "computer & electronics". "Computer & electronics" now is a high-level item replacing 2 low-level individual items (pages). All user activities on 2 individual pages will be aggregated into the "computer & electronics" item. The rows of the "user - item" matrix has millions of social users and its columns are those high-level items. The motivation to use "user - item" matrix to generate user preferences is based on the assumption that users may express similar sentiments on items with similar items. For example, a user who likes TV/movies is likely to like musician band page as well. In addition, the number of high-level items is significantly less than the number of low-level items. The number of items now is reduced to the number of categories which is much less than the number of individual pages. Therefore, using high-level items can solve the data

[1]http://mahout.apache.org/

sparse problem.

Another important issue in the item-based collaborative filtering is how to compute the similarity between items and then how to select the most similar items. The prediction is made by taking a weighted average of the target user's ratings on these similar items. The basic idea in similarity computation between item i and item j is to first isolate the users who have rated both of these items and then to apply a similarity computation to determine the similarity $s_{i,j}$. There are many different ways to compute the similarity. Here we describe 5 measures. We also investigate which measure gives the minimum approximation error.

- Cosine similarity: The similarity is measured by computing the cosine of the angle between two vectors. Similarity between item i and item j, denoted by $sim(i, j)$ is defined by $sim(i, j) = cos(\vec{i}, \vec{j}) = \frac{\vec{i} \bullet \vec{j}}{\|\vec{i}\| * \|\vec{j}\|}$, where "•" denotes the dot-product of the two vectors.
- Pearson correlation similarity: Similarity between items i and j is measured by computing the Pearson correlation $corr_{i,j}$. Let U denote the set of users who both rated i and j, then the correlation similarity is defined by $corr_{i,j} = \frac{\sum_{u \in U}(R_{u,i} - \vec{R}_i)(R_{u,j} - \vec{R}_j)}{\sqrt{\sum_{u \in U}(R_{u,i} - \vec{R}_i)^2}\sqrt{\sum_{u \in U}(R_{u,j} - \vec{R}_j)^2}}$, where $R_{u,i}$ denotes the rating of user u on item i, $\vec{R}_i$ is the average rating of the i-th item.
- Tanimoto coefficient similarity: A "similarity ratio" is given over bitmaps, where each bit of a fixed-size array represents the presence or absence of a characteristic in the plane being modeled. The definition of the ratio is the number of common bits, divided by the number of bits set in either sample. Presented in mathematical terms, if item X and Y are bitmaps, X_i is the i-th bit of X, and $\wedge$, $\vee$ are bitwise and, or operators respectively, the the similarity ratio T_s is $T_s(X, Y) = \frac{\sum_i (X_i \vee Y_i)}{\sum_i (X_i \wedge Y_i)}$. If each item is modeled instead as a set of attributes, this value is equal to the Jaccard Coefficient of the two sets. Tanimoto goes on to define a distance coefficient based on this ratio, defined over values with non-zero similarity $T_d(X, Y) = -log_2(T_s(X, Y))$.
- Log-likelihood based similarity: It is similar to the Tanimoto coefficient-based similarity, though more difficult to understand intuitively. It is a metric that does not take account of individual preference values. It is also based on the number of common items between two users, but, its value is more an expression of how unlikely it is for two users to have so much overlap, given the total number of items out there and the number of items each user has a preference for.
- Euclidean distance based similarity: It is most commonly used distance between two points (two vectors i and j). It is defined as $sim(i, j) = \sqrt{\sum_{n=1}^{N}(i_n - j_n)^2}$, where N is the dimension of vectors (the number of users in our case).

We describe two evaluation metrics for choosing best similarity measure using item-based collaborative filtering (MAE and RMSE). The results will be described in Section V.

(1) *Mean Absolute Error (MAE)* between actual user ratings (X_i) and predictions (Y_i) is a widely used metric. MAE is a measure of the deviation of recommendations from their true user-specified values. For each pair of $< X_i, Y_i >$, this metric considers the absolute error: $| X_i - Y_i |$. The MAE is computed by first summing these absolute errors of the N corresponding rating - prediction pairs and then computing the average. Mathematically, $MAE = \frac{\sum_i^N |X_i - Y_i|}{N}$. The lower the MAE, the more accurately CF predicts the user preferences.

(2) *Root Mean Squared Error (RMSE)* is also a frequently used measure of the differences between the actual user ratings (X_i) and the predicted user preferences (Y_i). For each pair of $< X_i, Y_i >$, the RMSE is defined as: $RMSE(X, Y)$ = $\sqrt{\frac{\sum_{i=1}^{N}(X_i - Y_i)^2}{N}}$. The lower the RMSE, the more accurately CF predicts user preferences.

Weighted sum strategy: We use the weighted sum strategy to approximate the user preference on each item. It computes the predicted value (preference score) $P_{u,i}$ on an item i for a user u by computing the sum of the ratings given by the user on the items similar to i. Each rating is weighted by the corresponding similarity $S_{i,j}$ between items i and j. It is formally defined as:

$$P_{u,i} = \frac{\sum_{j \in IS} S_{i,j} * R_{u,j}}{\sum_{j \in IS} |S_{i,j}|},$$

where:
IS: the set of all items similar to i;
$\mathbf{R}_{u,j}$: the rating on item j from user u, which is the normalized number of "likes" or normalized "rating" score from user u on item j.

B. PeerInf: Peer Influence

In social psychology, herd behaviors may occur frequently in everyday decisions based on learning from others' information [24]. In social platforms, it is also very common to see that content generated by one person could be viewed, cited, or duplicated by others. For example, customer reviews on a product written by the first reviewer usually has higher credibility than others and has a big impact on future review content. In this paper, we incorporate this herding social effect in predicting sentiments of comments.

We assume that if most of previous comments in one discussion are positive, it is likely to give a positive comment, and similarly for the negative case. If it is the first comment, then the corresponding sentiment is mostly decided by commentor's own opinion. For example, we randomly pick $1,000$ posts from 5 different Facebook pages and $1,000$ discussion threads from 5 different airlines on the Flyertalk.com forum. The average number of comments per post and per thread is 794 and 32, respectively. The sentiments are identified by the state-of-the-art textual algorithm first and then manually corrected. The Figure 2 shows that most of comments have similar sentiment orientations. There are 900+ Facebook posts and 850+ forum threads with $80 - 90\%$ percentage of same sentiments of their posts. In addition, each discussion thread

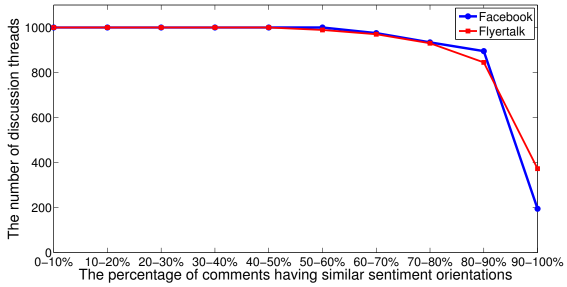


Fig. 2: Most of people have similar sentiment orientations within the same discussion on Facebook and Flyertalk.

has multiple pages of comments if it has more than a specific number of comments. For example, each page has maximum 50 comments on Facebook by default. Usually, it's not likely for users to read them all from previous pages when they make their own ones. Therefore, we introduce an exponential decay factor which weakens the effect of comments on higher page numbers. The older the page, the smaller the impact of its sentiments on the current user. For the i^{th} page, the weight is e^{1-i}. The current page is the first page. The oldest one is in the last page. We name this feature "PeerInf" in our system. It is mathematically defined as follows.

$$\text{PeerInf} = \mathbb{1}_{\{NF\}} * \frac{\sum_i PosR_i * e^{1-i}}{N} \qquad (\star)$$

where:

$$\mathbb{1}_{\{NF\}} = \begin{cases} 1 & \text{if it is not the first comment;} \\ 0 & \text{otherwise.} \end{cases}$$

i: the page number;
N: total number of pages;
$\mathbf{PosR}_i$: the positive ratio of all previous comments in the i^{th} page.

C. GenCat: User Profiles

Furthermore, the types of topics (e.g., Facebook/Twitter page category, product category on Amazon.com, airlines on Flyertalk.com) and the user profile information (e.g., gender) are also weak indictors of sentiment orientation. Due to the privacy setting, we could not get other user profile information, such as geo-location, user friends, education backgrounds, user interests, etc. We randomly pick 3 different page categories from Facebook and Twitter (politician, technology, and fashion) to see their sentiment distribution in terms of gender and topic type (category). Table I shows that female are more positive than male and fashion page has a higher percentage of positive sentiments than politician page on Facebook and Twitter. We also checked some other categories and found that they follow the same pattern. We name this feature "GenCat" in our system.

D. TextSent: Textual Sentiment

Besides the user social effects, the text itself also plays an important role in sentiment identification. Therefore, we

TABLE I: Differences of sentiment orientations under different genders and topics (e.g., page categories) on Facebook and Twitter.

Name (Category)	Gender	Positive ratio	Total number of comments+tweets
Barack Obama (Politics)	M F	0.61 0.69	6,837,096
Chicago Bulls (Sports)	M F	0.68 0.79	462,092
DKNY (Fashion)	M F	0.94 0.96	14,284

here describe our textual sentiment algorithm. We consider three types of values: positive, negative, and objective. Our textual sentiment identification algorithm integrates the following three different individual components. The first is a rule-based method extended from the basic compositional semantic rules [22] which include twelve semantic rules and two compose functions. Take Rule A for example: If a sentence contains the key word "but", then consider only the sentiment of the "but" clause. According to this rule, the following statement is considered positive: *"I've never liked that director, but I loved this movie."*. Compose functions generate integers from -5 to $+5$ as output to represent sentiment scores. The second component is a frequency-based method. We argue that the sentiment should not be simply classified as positive, negative, or objective but a continuous numerical score (e.g., -5 to $+5$) to reflect the sentiment strength. The strength of a sentiment is expressed by the adjective and adverb used in the sentence. We consider two kinds of phrases that derive numerical scores: the phrases in the forms of adverb-adjective-noun (abbreviated as AAN) and verb-adverb (VA). The scores of key words were used are calculated based on a large collection of customer reviews, each of which is associated with a rating. The details of score calculation can be found in our previous work [26]. Here, we present a few examples. "Easy" has a score of 4.1, "best" 5.0, "never" -2.0, and "a bit" 0.03. Furthermore, the third bag-of-word component considers special characters commonly used in social media text, such as emoticons, negation words and their corresponding positions, and domain-specific words. For example, ':)' is a positive sentiment and ':(' a negative sentiment. Some Internet language expresses positive opinions like "1st!", "Thank you, Obama", "Go bulls!". Some domain specific words are also included, like "Yum, Yummy" for food related brands. Finally, a random forest machine learning model is applied to the features generated from outputs of the three components. The outputs are represented as three basic features (TS_1, TS_2, TS_3) and two derived features (TS_1+TS_2, TS_1-TS_2). Our sentiment identification algorithm is trained on manually labeled Facebook comments and Twitter tweets and achieve an accuracy of 86%. We name this textual sentiment result as feature "TextSent" in our system.

V. Experiments and Results

In this section, we first describe four datasets. Data can be downloaded through the Facebook Graph API, Twitter

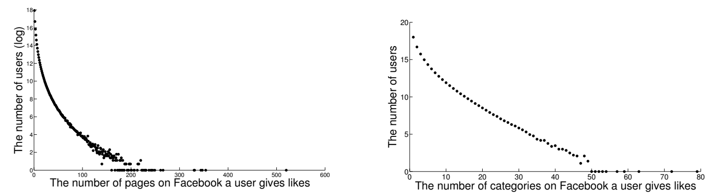
Search/Streaming API, Amazon.com Product API, and Flyertalk.com RSS feed. After collecting data, we design a filtering system to remove spam users. Once the data is clean, for each comment or review, we run the textual algorithm to get the textual sentiment result ("TextSent"). We run distributed item-based collaborative filtering on the "user-item" matrix to get the user preference for each item ("UserPref"). Then, peer influences ("PeerInf") is calculated based on the definition (Eq. ($\star$)). Some other additional information, such as user gender, topics, come along with the data. We train and test our model on Facebook comments, Twitter tweets, product reviews, and Flyertalk comments. All these data were manually labeled as either positive, negative or neutral.

A. *Experimental Data and Setup*

For each page in Facebook data, we download all posts, comments, likes, and corresponding user profile since the first day they opened their account on Facebook. For each Twitter page, we download tweets, follower, and corresponding user profile. For each product on Amazon.com, we download all product information and corresponding reviews. These products are from 10 randomly picked categories. For the Flyertalk.com, we use all discussions from July, 2010 to June, 2012 via their RSS feed.

Due to differences among these datasets, some features in our system would be varied or missed. For example, the element M_{ij} in the "user-item" matrix M for Facebook is from the number of "likes" given by the user i to the category j. However, M_{ij} is from the 5-star rating score for product reviews, generated from "if following or not" on Twitter. And this matrix is missed for Fyertalk.com forum data because there is no other quantitative ways to represent user's preferences on different airlines except their discussion comments. The Table II lists the features we use to do training on each dataset. Some features are missed because of the lack of data.

Data Cleaning: Due to the data redundancy and noises, we designed some rules to clean our data. We take Facebook for example. By June 1, 2012, we had $12,063$ pages and approximately 280 million users in our database. We sort Facebook pages by the number of fans. As the first step, we removed pages which have very few fans or the number of post likes, because fewer user post likes can not contribute to statistically significant results. There were $7,580$ Facebook pages left. Further, we designed a simple and conservative strategy to filter out spam users. We found that on average, a user becomes a fan of, and likes posts on 8.99 categories and 7.45 pages, as shown in the Figure 3a and the Figure 3b. Users connecting to an extremely large number of pages are likely to be spam users or bots. For example, there is one user who likes posts across 520 different pages. Most of the users are interested in a handful of pages. In our experiments, we set the threshold of 150 to discard users like posts on more than 150 pages. In addition, we also observed that some spam users like many posts launched by the same page. For example, there is one user who liked $7,963$ posts out of total $8,549$ posts. We set this threshold to be 90% for every user except the page owner. Lastly, we also removed users who posted many duplicated comments on the same brand and most of the duplicated comments contain URL links. A test on Barack Obama's page, found $209,864$ duplicated comments out of $2,987,505$ in total. After these data cleaning steps, the data size was reduced significantly. Table III shows the stats of cleaned data for computing "UserPref" value for each user on each category.



(a) Low-level: individual page (b) High-level: category

Fig. 3: The distribution of individual pages and categories on Facebook access by users. The Y-axis is taken by log function.

TABLE III: The stats of our experimental data for calculating user preference.

	Facebook	Twitter	Amazon.com
# of unique users	$97,699,832$	55,367,807	$6,357$
# of items (category)	150	150	10

Experimental Setup: The most time consuming part of our experiments is user preference approximation using item-based collaborative filtering. We used a Hadoop based system which contains 20 Linux-based nodes. Each node is a PC with AMD dual cores and 2GB of RAM.

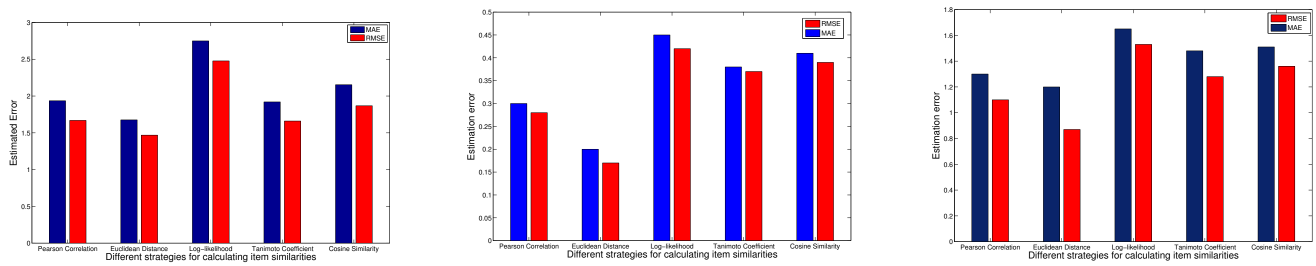
B. *Experimental Results*

In this section, we first present the minimum estimation error for 5 different similarity measures of using the item-based collaborative filtering to estimate user preferences. And we also compare the time taken by Hadoop-based and single-node based collaborative filtering algorithms. Then, we show the sentiment classification accuracy and how well each feature contributes to the prediction accuracy.

1) User Preference Approximation: Figure 4 shows MAE and RMSE under different similarity measures of the user preference approximation using item-based collaborative filtering engine. The Euclidean distance measurement gives the best performance for both MAE and RMSE. The smaller the error, the better the performance of a similarity measure. The running time of user preference prediction was significantly improved using distributed computing. Hadoop implementation took about 34 and 21 minutes on average for Facebook and Twitter, whereas the single CPU implementation can not complete even in 10 hours. For the small data from Amazon.com, it is finished very quickly. We do not calculate user preference for Flyertalk because there is no user rating functionality on Flyertalk.

TABLE II: The features of learning model for 4 datasets and their differences. Topic is modified based on the raw Facebook category. The total categories we used in our experiments is 150. "$\times$": missed; "$\surd$": existing.

Data Source	TextSent	UserPref	PeerInf	GenCat	
				Gender	Topic
Facebook	comments	user-post likes on categories	$\surd$	$\surd$	predefined category
Twitter	tweets	user-category following	$\surd$	$\surd$	predefined category
Amazon	product reviews	user-product rating	$\surd$	$\times$	product category
Flyertalk	airline discussion comments	$\times$	$\surd$	$\times$	airline types



(a) Facebook data

(b) Twitter data

(c) Amazon.com data

Fig. 4: MAE and RSME of item-based collaborative filtering under 5 different similarity strategies.

2) Sentiment Classification: We now present the classification accuracy for sentiment prediction and the contribution of each factor ("TextSent", "UserPref", "PeerInf", "GenCat") to the prediction accuracy. We chose $4,000$ subjective instances ($2,000$ positive and $2,000$ non-positive) in each data source to form our 4 training sets. We run support vector machine (SVM) with 10-fold cross validation. Table IV shows the classification accuracy of using compositional semantic rules (CSR) algorithm [22], supervised and unsupervised methods [28], some previous methods, and the accuracy with different combinations of features from our proposed user social effects. The results show that using all four major features gives the best performance, with each feature incrementally enhancing the prediction accuracy. This means that incorporating social effects such as user preference, peer influence, user profile, and topics, is potentially helpful to improve the sentiment identification for social media data. For the short texts, such as Facebook comments and Twitter tweets, user preference ("UserPref") get higher improvement than peer influence ("PeerInf") and user profile information ("GenCat"). Because "likes" or ratings carry more opinions than comments. However, the peer influence has a higher impact of sentiment identification on longer texts such as products reviews on Amazon.com and flight discussions on Flyertalk. Most of long texts deliver much more comprehensive information on which future reviewers or commentors rely more. And it is more difficult for future reviewers or commentors to read through many old reviews or comments. Due to the missing user profile information and missing functionalities on Flyertalk, the accuracy only changes when peer influence is added.

TABLE IV: The classification accuracy under different combinations of features and using existing sentiment algorithms. SS: semantic features + syntactic features used in [28].

Baselines	**Sentiment Accuracy**			
	Facebook	Twitter	Amazon.com	Flyertalk
Compositional Semantic Rules (CSR)	72.0%	74.0%	57.0%	63.0%
SVM on SS	60.0%	60.7%	63.5%	63.2%
Logistic Regression on SS	61.2%	61.2%	63.7%	62.4%
HMM on SS	61.0%	61.7%	64.0%	64.2%
CRF on SS	61.4%	60.5%	69.3%	66.5%
Pang's method [7]	84.3%	81.0%	82.9%	69.7%
Liu's method [2]	84.8%	85.0%	74.9%	66.6%
Model Features	**Sentiment Accuracy**			
	Facebook	Twitter	Amazon.com	Flyertalk
TextSent only	86.0%	86.0%	74.0%	71.0%
TextSent + UserPref	89.4%	89.2%	79.1%	71.0%
TextSent + UserPref + PeerInf	91.7%	92.0%	84.2%	75.5%
TextSent + UserPref + PeerInf+ GenCat	**92.6%**	**93.5%**	**84.2%**	**75.5%**

VI. Conclusion and Future Work

In this work, we proposed a systematic framework to classify sentiments for social media data. Various types of user social effects have been incorporated into our system including user preference, peer influence, user gender, and topics. We also presented a distributed item-based collaborative filtering technique to approximate user preferences. We model how previous comments or reviews affect sentiment of current comments. We conducted our experiments on four data from Facebook.com, Twitter.com, Amazon.com, and Flyertalk.com. The experimental results demonstrated that we can get significantly better classification accuracy compared to textual sentiment identification algorithms and state-of-the-art supervised and unsupervised learning algorithms.

In the future, we will incorporate more network data, such as friendship, geo-location, and so on. Although more complicated model could make the model more realistic, the computational complexity will become more challenging. In addition, finding a more efficient collaborative filtering algorithm and sampling technique would be extremely helpful as well.

References

[1] B. J. Jansen and M. Zhang and K. Sobel and A. Chowdury, *Twitter power: Tweets as electronic word of mouth*, J. Am. Soc. Inf. Sci. Technol. New York, USA: John Wiley & Sons, Inc., 2009.

[2] M. Hu and B. Liu, *Mining and summarizing customer reviews*, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. New York, USA: ACM, 2004.

[3] S. Asur and B. A. Huberman, *Predicting the Future with Social Media*, Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01. Washington, DC, USA: ACM, 2010.

[4] M. Joshi and D. Das and K. Gimpel and N. A. Smith, *Movie reviews and revenues: an experiment in text regression*, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, USA: ACL, 2010.

[5] L. Zhang and B. Liu, *Identifying noun product features that imply opinions*, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2. Portland, USA: ACL, 2011.

[6] R. Narayanan and B. Liu and A. Choudhary, *Sentiment analysis of conditional sentences*, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1. Singapore: ACL, 2009.

[7] B. Pang and L. Lee and S. Vaithyanathan, *Thumbs up?: sentiment classification using machine learning techniques*, Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10. ACL, 2002.

[8] P. D. Turney, *Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews*, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, USA: ACL, 2002.

[9] J. Liu and S. Seneff, *Review sentiment scoring via a parse-and-paraphrase paradigm*, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1. Singapore: ACL, 2009.

[10] M. Ryan and H. Kerry and N. Tyler and W. Mike and R. Jeff, *Structured Models for Fine-to-Coarse Sentiment Analysis*, Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. ACL, 2007.

[11] J. M. Wiebe and R. F. Bruce and T. O'Hara and P. Thomas, *Development and use of a gold-standard data set for subjectivity classifications*, Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. College Park, USA: ACL, 1999.

[12] J. Wiebe and T. Wilson, *Learning to disambiguate potentially subjective expressions*, Proceedings of the 6th conference on Natural language learning - Volume 20. ACL, 2002.

[13] H. Yu and V. Hatzivassiloglou, *Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences*, Proceedings of the 2003 conference on Empirical methods in natural language processing. ACL, 2003.

[14] T. Wilson and J. Wiebe and R. Hwa, *Just how mad are you? finding strong and weak opinion clauses*, Proceedings of the 19th national conference on Artificial intelligence. San Jose, USA: AAAI, 2004.

[15] S. Kim and E.Hovy, *Determining the sentiment of opinions*, Proceedings of the 20th international conference on Computational Linguistics. Geneva, Switzerland: ACL, 2004.

[16] E. Riloff and J. Wiebe, *Learning extraction patterns for subjective expressions*, Proceedings of the 2003 conference on Empirical methods in natural language processing. ACL, 2003.

[17] V. Stoyanov and C. Cardie, *Topic identification for fine-grained opinion analysis*, Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1. Manchester, United Kingdom: ACL, 2008.

[18] A. Popescu and O. Etzioni, *Extracting product features and opinions from reviews*, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, Canada: ACL, 2005.

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

[20] L. Ku and Y. Liang and H. Chen, *Opinion Extraction, Summarization and Tracking in News and Blog Corpora*, AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. AAAI, 2006.

[21] X. Ding and B. Liu and P. S. Yu, *A holistic lexicon-based approach to opinion mining*, Proceedings of the 2008 International Conference on Web Search and Data Mining. Palo Alto, USA: ACM, 2008.

[22] Y. Choi and C. Cardie, *Learning with compositional semantics as structural inference for subsentential sentiment analysis*, Proceedings of the Conference on Empirical Methods in Natural Language Processing. Honolulu, USA: ACM, 2008.

[23] C. Tan and L. Lee and J. Tang and L. Jiang and M. Zhou and P. Li, *User-level sentiment analysis incorporating social networks*, Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, USA: ACM, 2011.

[24] A. .V. Banerjee, *A Simple Model of Herd Behavior*, The Quarterly Journal of Economics. , 1992.

[25] K. Zhang and R. Narayanan and A. Choudhary, *Voice of the customers: mining online customer reviews for product feature-based ranking*, Proceedings of the 3rd conference on Online social networks. Boston, USA: ACM, 2010.

[26] K. Zhang and Y. Cheng and Y. Xie and D. Honbo and A. Agrawal and D. Palsetia and K. Lee and W. Liao and A. Choudhary, *SES: Sentiment Elicitation System for Social Media Data*, ICDM Workshops. IEEE, 2011.

[27] K. Zhang and R. Narayanan and A. Choudhary, *Item-based collaborative filtering recommendation algorithms*, Proceedings of the 10th international conference on World Wide Web. Hong Kong: ACM, 2001.

[28] K. Zhang and Y. Xie and Y. Cheng and D. Honbo and D. Doug and A. Agrawal and W. Liao and A. Choudhary, *Sentiment identification by incorporating syntax, semantics and context information*, Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. Portland, USA: ACM, 2012.

[29] K. Dave and S. Lawrence and D. Pennock, *Mining the peanut gallery: opinion extraction and semantic classification of product reviews*, Proceedings of the 12th international conference on World Wide Web. Budapest, Hungary: ACM, 2003.

[30] Y. Zhang and J. Tang and J. Sun and Y. Chen and J. Rao, *MoodCast: Emotion Prediction via Dynamic Continuous Factor Graph Model*, Proceedings of the 10th IEEE International Conference on Data Mining. Sydney, Australia: IEEE, 2010.

[31] J. Tang and J. Sun and J. Rao and Y. Chen and A. Fong, *Quantitative Study of Individual Emotional States in Social Networks*, IEEE Transactions on Affective Computing. IEEE, 2012.