1. INTRODUCTION

Nowadays decision process is based on all the personal information available from online textual reviews. For example, the customer will decide what to buy if he or she sees valuable reviews posted by others, especially user’s trusted friend. Based on the idea that a high star rating of product is usually related to the good reviews, here this idea helps both review and reviewers to predict the ratings of a particular product. Hence, how to mine reviews and the relation between reviewers in social networks has become an important issue in web mining, machine learning and natural language processing.

More focus is on the rating prediction task. However, user’s rating information is not always available on many review websites. But reviews containing detailed product, user opinion information has a great reference value for a user’s decision. Most important of all, a given user on website is not possible to rate every item therefore many unrated items are left behind. It is inevitable in many rating prediction approaches e.g. [1], [4]. Review/comment, as we all know, is always available. In such case, it’s convenient and necessary to influence user reviews to help predicting the unrated items. Generally, user’s interest is stable in short term, so user topics from reviews can be representative. Consider an example, in the category of Health care band, Digital thermometer different people have different requirements.

Sentiment analysis is the most fundamental and important work in extracting user’s interest preferences. Sentiment basically evaluates or determines user’s own attitude on items. Practically, it is more important to provide numerical scores rather than binary decisions. Generally, reviews are divided into groups such as positive and negative. However, it is difficult for customers to make a choice when all candidate products reflect positive sentiment or negative sentiment.

To purchase particular product customers need to know how good the product is. It’s also agreed that different people may have different sentimental expression preferences. For example, some users prefer to use “nice” to describe an “excellent” product, while other users may prefer to use “good” to describe a “just so so” product.

In our daily life, customers are mostly preferred to buy products with highly- praised reviews. That is, customers are more concerned about item’s reputation. To obtain the reputation of a product, sentiment in reviews is necessary. To a given product, if we know user sentiment, we can infer the reputation.

When we search the net for purchasing, both positive and negative reviews are valuable. Positive reviews express the advantages of a product and on the other hand, negative reviews can be used obtain the shortcomings of product in case of being cheated. We observe that reviewers’ sentiment will influence others: if a reviewer has positive or negative sentiment (i.e. clear like or dislike), other users will pay much attention to him/her. However, it is difficult for reviewers to take decision if review sentiments are neutral.

We use a sentiment-based rating prediction method. In our work, we make use of social users’ sentiment to infer ratings. Fig. 1 is an example that illustrates motivation of our work. Firstly, we extract product features from reviews. Then, we identify the sentiment words, which describe the product features by eliminating “stop” words and “noise” words. Besides that, we also use sentiment dictionaries to calculate sentiment polarities of a specific user on an item/product. In Fig.1, the last user is interested in product features such as quality, price and brand, so based on the reviews and

**Abstract**— In recent years, we have witnessed numerous reviews from different review websites. These review websites give a great opportunity to share our viewpoints for various products we purchase. Due to loads of information, we face problem in mining valuable information from reviews to understand a user’s preferences and make an accurate recommendation. Traditional recommender systems (RS) consider some factors, such as product category, user’s purchase records and geographic location. In this work, we propose a sentiment-based rating prediction method (RPS) to improve prediction accuracy in recommender systems along with Dynamic question list for review generation. We use a social user sentimental measurement approach to calculate user’s sentiment on items/products along with that we also consider user’s interpersonal sentimental influence and product reputation, which can be inferred from user’s reviews. At last, we combine three factors-user sentiment similarity, item reputation similarity and interpersonal sentimental influence into our recommender system to make an accurate rating prediction. We also generate Dynamic question list for easy review generation.

**Keywords**— Item reputation; Reviews; Rating prediction; Recommender system; Sentiment influence; User sentiment; HDFS (Hadoop Distributed File System); HIVE; RPC(Remote Procedure Calls); ACLs(ACCESS Control Lists)
the sentiment dictionaries, the last item will be recommended and a dynamic question list will get generated based on the selected product for further review generation. Compared with \[5\], \[11\], \[16\], \[26\], \[27\], the main difference between our system and those systems is that: their work mainly focuses on classifying users into binary sentiment (i.e. positive or negative). In our paper, we mine social user’s sentiment, as well as we explore interpersonal sentimental influence and item’s reputation. Finally, combine all of them in the recommender system.

The main contributions of our approach are as follows: 1) we make use of users sentiment for rating prediction. User sentiment influence shows how the sentiment spreads among the trusted users. Item reputation similarity shows the relevance of items. 2) We combine the three factors: user sentiment similarity, item reputation similarity, and interpersonal sentimental influence into a probabilistic matrix factorization framework for accurate rating prediction.

The remainder of this paper is as follows: In Section 2, we present the related work about rating prediction. In Section 3, the proposed sentiment-based rating prediction method is described thoroughly. In Section 4, the System architecture is described. Conclusions are drawn in Section 5.

2. RELATED WORK

In this section, we survey recent work related to our approach. We review some approaches based on collaborative filtering (CF). Also, review based approaches as well as the sentiment based applications are provided in detail.

A. Collaborative Filtering

Collaborative filtering is used to predict user preferences for unrated products, after which recommendations can be provided to users from list of most preferred items.

To improve recommendation performance, many CF algorithms have been proposed \[12\],[13],[15]. One of the most well-known CF algorithms is the user-based CF algorithm proposed in [15]. The basic idea is that people Expressed similar preferences in the past will prefer to buy similar items in the future. Item-based CF algorithm [12] produces the rating from a user to an item based on the average ratings of similar or correlated items by the same user. It obtains better performance in computing the similarity between items. Fletcher et al. [13] propose a CF-based service recommendation method that considers users’ personalized preferences on nonfunctional attributes.

B. Matrix Factorization

Matrix factorization is an approach for low-dimensional matrix decomposition (Basic MF [1]). Some matrix factorization based social recommendations are proposed to solve the “cold start” problems. Jamali et al. [4] explore a matrix factorization based approach for recommendation in social networks. They incorporate the mechanism of trust propagation into the recommendation model. Jiang et al. [3] propose another important factor, the individual preference. They conduct experiments on Renren dataset and Tencent Weibo dataset in China, and the results demonstrate the significance of social contextual factors (individual preference and interpersonal influence) in their model. Qian et al. [6] propose a personalized recommender model (PRM) combing with user interpersonal interest similarity, interpersonal influence and personal interest factor. They make use of categories of products, and user personal interest is the main contributions.

C. Reviews based Applications

There are also many reviews based work for the task of recommendation. Qu et al. [14] propose a bag-of-opinions model to predict a user’s numeric rating in a product review. Wang et al. [10] propose a review rating prediction method by incorporating the social relations of a reviewer. In addition, they classify the social relations of reviewers into strong social relation and ordinary social relation. Zhang et al. [17] incorporate various product review factors including content related to product quality, time of the review and older positive customer reviews. They have used a product ranking model to calculate the ranking score. Luo et al. [18] define and solve a new problem: aspect identification and rating, together with overall rating prediction in unrated reviews. They propose a personalized recommender model (PRM) combing with user interpersonal interest similarity, interpersonal influence and personal interest factor. They make use of categories of products, and user personal interest is the main contributions.

D. Sentiment Based Applications

Sentiment analysis can be conducted on review-level, sentence-level, and phrase-level. Review-level analysis [19], [20] and sentence-level analysis [21] attempt to classify the sentiment of a review to one of the sentiment polarities including positive, negative and sometimes neutral. While phrase-level analysis [26], [22] attempt to extract the sentiment polarity of each feature that expresses users attitude to the specific feature of a specific product. The
main task of phrase-level sentiment analysis is the construction of sentiment lexicon. Lu et al. [9] propose an optimization framework that provides a unified and principled way to combine different sources of information for learning a context-dependent sentiment lexicon. There are many approaches leveraging sentiment analysis for personalized recommendation [8], [24], [26], [27]. Zhang et al. [8] propose a self-supervised and lexicon-based sentiment classification approach to determine sentiment polarity of a review containing textual words as well as emoticons. And they use sentiment for recommendation. Lee et al. [24] propose a recommender system that uses concept of Experts to find relevant recommendations. Lei et al. [27] leverage phrase-level sentiment analysis to infer a specific item’s reputation. Zhang et al. [26] propose an Explicit Factor Model (EFM) to generate an explainable recommendation they extract explicit product features and user opinions by phrase-level sentiment analysis on reviews.

E. Security of Data in HDFS

Initially, Hadoop didn’t authenticate users or services, and there was no data privacy. As Hadoop is mainly a platform for data analytics and processing, security issues need to be solved. Hadoop has developed certain frameworks or methods to achieve this security. They are as follows:

- Mutual Authentication with Kerberos RPC on RPC connections– SASL/GSSAPI was used to implement Kerberos. It also mutually authenticates users, processes, and Hadoop services on RPC connections.
- Enforcement of HDFS file permissions – Access control to files in HDFS could be enforced on file permissions by Name Node using Access Control Lists (ACLs) of users and groups.
- Block Access Tokens for Access Control to Data Block- When access to data blocks is needed, then Name Node makes an access control decision based on HDFS file permissions and issues Block access tokens that could be sent to the Data Node for block access requests.
- Network Encryption - This includes connections using Kerberos RPC and subsequent authentication. Web consoles and Map reduce shuffle operations can be encrypted by configuring them to use SSL. Finally, HDFS File Transfer can also be configured for encryption.

3. PROPOSED SYSTEM

The purpose of our approach is to find sentiments of users from reviews and predict social users’ ratings.

3.1 SYSTEM ARCHITECTURE

Fig.2, describes System architecture/flow for our proposed system. There are huge amount of textual reviews generated each and every day on online websites such as Amazon, Flipkart etc. So, hereby to analyze few predictions on different products we consider Amazon Health care online reviews data further to evaluate the ratings and for product recommendation. Later, this structured data is pushed onto HDFS (Hadoop Distributed File System) so that data can be processed further by data mining technique that is Map reduce. This processing results into 3 judgements as Positive, Negative, Neutral. This processed data is stored onto HDFS later for data verification and the verified data will get retrieved with the help of HIVE from HDFS this way a final report is generated based on the above process and a Sentiment Report is stored onto HDFS. Hence, the Sentiment Report generated helps in product rating and recommendations to the user and thus results in fulfilling our objective.

3.2 WORKING OF PROPOSED SYSTEM

In this paper, we firstly extract product features from user review corpus, and then we introduce the method of identifying social users’ sentiment. In addition, we describe the three sentimental factors. At last we fuse all of them into our sentiment-based rating prediction method (RPS).

A. Extracting Product Features

Product features mainly focus on the discussed issues of a product. In this paper, we extract product features from textual reviews using LDA [7]. We mainly want to get the product features including some named entities and some product/item/service attributes. LDA is a Bayesian model, which is utilized to model the relationship of reviews, topics and words. To construct the vocabulary, we firstly consider each user’s review as a collection of words without considering the order. Then we filter out “Stop Words”, “Noise Words” and
sentiment words. Some prepositions, articles, and pronouns are considered as “Stop Words”. For extracting product features, we need to filter the noisy features from the candidate set. Users in different topics care about different product features.

B. User Sentimental Measurement

In our paper, we have created Sentiment Dictionary to calculate social user’s sentiment on items. We have created positive sentiment words list and named it as POS -Words; also, the negative sentiment words list and named it as NEG- Words. Our sentiment dictionary (SD) includes 2006 POS- Words and 4783 NEG-Words. We firstly divide the original review into several clauses by the punctuation mark. Then for each clause, we look up the dictionary SD to find the sentiment words before the product features. A positive word is initially assigned with the score +1.0, while a negative word is assigned with the score -1.0.

To improve the precision of sentiment mapping, we add two main linguistic rules as: By applying conjunctive rules [9], [23], [25], “and” conjunctives between clauses generally implies same sentiment polarity. Other “and”-like terms are: as well as, likewise. “but” conjunctives between clauses usually expresses the opposite sentiment polarity. Similar terms like “but” are: however, nevertheless, though, and etc. Distinguish between product features and sentiment words. Some features (i.e. noun) like “noise”, “mistake” are with clear negative sentiment polarity, while “pleasure”, “happiness” are with clear positive sentiment polarity. Here we treat these words as sentiment words and collect them into sentiment dictionary (SD). The words like “acclamation”, “pleasure”, and “happiness” will be collected into POS -words of SD, the words like “noise”, “stink”, and “mistake” will be collected into NEG-words of SD. When deciding the sentiment score of such a phrase (e.g. “noise”) in a review, we will give a score of -1.0.

C. Three Sentimental Factors

This section describes major components of our approach. Each sentiment factor is described as follows:

User Sentiment Similarity- Generally, user’s friends are trustworthy [2], [4], [6]. If a user has similar interest preferences with his/her friends, then he/she may hold similar attitudes towards the item. Based on this view, we firstly get all users’ sentiment, and then calculate the sentiment similarity between the user and his/her friends.

Interpersonal Sentiment Influence- When we search the inter-net for purchasing, we are more concerned with those users who posted five-star reviews or critical reviews. Especially, the critical reviews can reflect the deficiency of a product. In this case, we observe that reviewers’ sentiment will influence others, if a reviewer expressed clear like or dislike sentiment, other users will obtain the specific advantages or weaknesses about a product. However, the middle evaluations have little useful information.

In our paper, we argue that if a user always has explicit attitude about a product, his/her reviews will has a great reference value to others, and this user has a big influence on others. While a user always has neutral attitude will has a small reference value to others, and this user will has a small influence on others. Item Reputation Similarity- From typical item- based collaborative filtering algorithms in [12], we know that similar items can help in predicting ratings. Thus, it is important for us to find items that have similar features. In our work, we assume item’s reputation can indirectly reflect its real ratings. We leverage users’ sentiment distribution to infer item’s reputation. Based on users’ sentiment, we believe that if two items have similar sentiment distribution, then they may have similar reputation, and they will be posted with similar ratings.

D. Sentiment Based Reccomender Model

After taking the three sentimental factors above into consideration, we have three important constrain terms in our rating prediction model. They are: 1) Normalized user sentiment similarity 2) Normalized interpersonal sentiment influence. 3) Normalized item reputation similarity. According to the matrix factorization, we fuse the three factors into objective function.

E. Generation of Dynamic Question List

In our work, after prediction of ratings for products, most reviewed/top rated products are collaborated into a set and Dynamic question list is generated based on extracted product features from reviews. Consider an example of camera in which product features can be Brand, pixel quality, appearance and other additional features like modes (Panorama or night).

F. Security of our Dataset in HDFS

In our system, we are storing both input and output data on HDFS. Hence, it is ensure to maintain security of this data. UNIX file permissions can be applied to HDFS but they are not sufficient when we need special permissions for user and group. HDFS ACLs gives the ability to specify fine-grained file permissions for specific named users or named groups, not just the file’s owner and group. To use ACLs, we first need to enable ACLs on Name Node and then file permissions can be added.

4. RESULTS

In this section, we have discussed results generated from implementation of the paper.
Fig. 3 shows an example of the dataset we have used and discussed earlier, i.e., Amazon Health Care data. This dataset is in JSON format which consists of key-value pairs. Actual review is present with key “reviewText”.

**Fig 4: Mapper Code part**

```java
public class SubMapper extends MapReduceBase implements Mapper<InputSplit, Text, OutputCollector<Text, IntWritable>> {
    private Array.asList<String> listOfPositiveWord = null;
    private Array.asList<String> listOfNegativeWord = null;
    private Map<String, Integer> listOfActionSentimentLexicon = null;
    private JSONParser parser = null;

    public SubMapper() throws IOException, InterruptedException {
        super();
    }

    public void map(Text key, Text value, OutputCollector<Text, IntWritable> output, Reporter reporter) throws IOException {
        try {
            parser = new JSONParser();
        } catch (ParseException e) {
            e.printStackTrace();
        }

        JSONObject jsonObject = parser.parse(value.toString());
        JSONArray jsonArray = jsonNode.getJSONArray("reviewText");
        for (int i = 0; i < jsonArray.length(); i++) {
            String reviewText = jsonArray.getString(i);
            this.setActionSentiment(new ArrayList<String>(Arrays.asList(reviewText.split(" "))));
            output.collect(key, IntWritable.valueOf(1));
        }
    }
}
```

**Fig 5: Mapper Code part2**

Fig.4 and Fig.5 show code for mapper class in MapReduce framework. Fig. 4 consists of code to extract textual review from JSON input file which consists of Amazon health care data. Fig.5 consists of code to eliminate repeating character words from reviews. There can be various repeating character words which interprets the degree of sentiment. For example, I just love the product. In this example, word love is written as love which implies a higher positive sentiment degree.

In this paper, a recommendation model is proposed by mining sentiment information from social users’ reviews. We fuse user sentiment similarity, interpersonal sentiment influence, and item reputation similarity into a unified matrix factorization framework to achieve the rating prediction task. Besides that, we also generate Dynamic question list for review generation based on previous reviews for most reviewed products.

**REFERENCES**


