From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring

GUANGMING GUO, University of Science and Technology of China
FEIDA ZHU, Singapore Management University
ENHONG CHEN and QI LIU, University of Science and Technology of China
LE WU, Hefei University of Technology
CHU GUAN, University of Science and Technology of China

With the booming popularity of online social networks like Twitter and Weibo, online user footprints are accumulating rapidly on the social web. Simultaneously, the question of how to leverage the large-scale user-generated social media data for personal credit scoring comes into the sight of both researchers and practitioners. It has also become a topic of great importance and growing interest in the P2P lending industry. However, compared with traditional financial data, heterogeneous social data presents both opportunities and challenges for personal credit scoring. In this article, we seek a deep understanding of how to learn users' credit labels from social data in a comprehensive and efficient way. Particularly, we explore the social-data-based credit scoring problem under the micro-blogging setting for its open, simple, and real-time nature. To identify credit-related evidence hidden in social data, we choose to conduct an analytical and empirical study on a large-scale dataset from Weibo, the largest and most popular tweet-style website in China. Summarizing results from existing credit scoring literature, we first propose three social-data-based credit scoring principles as guidelines for in-depth exploration. In addition, we glean six credit-related insights arising from empirical observations of the testbed dataset. Based on the proposed principles and insights, we extract prediction features mainly from three categories of users' social data, including demographics, tweets, and networks. To harness this broad range of features, we put forward a two-tier stacking and boosting enhanced ensemble learning framework. Quantitative investigation of the extracted features shows that online social media data does have good potential in discriminating good credit users from bad. Furthermore, we perform experiments on the real-world Weibo dataset consisting of more than 7.3 million tweets and 200,000 users whose credit labels are known through our third-party partner. Experimental results show that (i) our approach achieves a roughly 0.625 AUC value with all the proposed social features as input, and (ii) our learning algorithm can outperform traditional credit scoring methods by as much as 17% for social-data-based personal credit scoring.

This research was partially supported by grants from the National Key Research and Development Program of China (Grant No. 2016YFB1000904), the National Science Foundation for Distinguished Young Scholars of China (Grant No. 61325010), the National Natural Science Foundation of China (Grant No. 61403358 and 61602147). This work was also partially supported by the Pinnacle Lab for Analytics @ Singapore Management University, and the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA).
Authors' addresses: G. Guo, E. Chen (corresponding author), Q. Liu, and C. Guan, School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, China; emails: {guogg, guanchu}@mail.ustc.edu.cn, {cheneh, qiliuql}@ustc.edu.cn; F. Zhu, School of Information Systems, Singapore Management University, Singapore, 178902; email: fdzhu@smu.edu.sg; L. Wu, School of Computer and Information, Hefei University of Technology, Hefei, Anhui 230009, China; email: lewu.ustc@gmail.com.

DOI: http://dx.doi.org/10.1145/2996465

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*; H.2.8 [**Database Management**]: Database Applications—*Data mining*

General Terms: Algorithms, Design, Experimentation

Additional Key Words and Phrases: Personal credit scoring, social data, consumer finance, P2P lending, features, user profiling

ACM Reference Format:
Guangming Guo, Feida Zhu, Enhong Chen, Qi Liu, Le Wu, and Chu Guan. 2016. From footprint to evidence: An exploratory study of mining social data for credit scoring. ACM Trans. Web 10, 4, Article 22 (December 2016), 38 pages.
DOI: http://dx.doi.org/10.1145/2996465

1. INTRODUCTION

The blossoming of social networking services has generated an unprecedented amount of social media data about individuals over time. Business and social insights attainable from the big social data are immensely valuable for a wide range of applications such as targeted marketing [Kempe et al. 2003], event detection [Sakaki et al. 2010], and stock market prediction [Bollen et al. 2011]. Recently, there has been tremendous interest in harnessing social media data for *personal credit scoring*, especially with the fast growth of a new business model called P2P lending [Bachmann et al. 2011; Zhao et al. 2016] in the online world. P2P lending, also known as social lending [Hulme and Wright 2006], or crowdfunding [Gerber and Hui 2013; Mollick 2014], refers to the Internet-based practice of lending money to unrelated or unfamiliar individuals. Although the online lending process is extremely efficient and time-saving, accurate credit checking for online applicants also becomes increasingly urgent for the P2P lending industry's development and prosperity.

Unfortunately, even state-of-the-art financial-data-based credit scoring systems are limited in meeting the huge demands of usable credit evaluations in the P2P lending industry. Challenges of traditional credit scoring methods for P2P lending industry mainly come from the following three aspects:

—**Data Coverage.** According to American Consumer Financial Protection Bureau,[1] about 1 in 10 American adults had no credit history (i.e., credit invisible) until 2015. Due to insufficient credit history, another 8% of American adults only have credit records that are "unscorable" by widely used credit scoring models, not to mention those in other less-developed countries. In a word, a large number of consumers are short of usable credit history.
—**Data Timeliness.** Unlike social data, financial transactions or credit records are not generated frequently and may become out-of-date for personal credit evaluation. If unexpected accidents happen to a given applicant, traditional credit risk management systems cannot alert lenders in time. Due to slow response to unexpected credit risks, it is estimated that the number of P2P lending companies in China will probably drop from 2,000+ to a couple hundred in the next few years.[2]
—**Data Availability.** Even for users with sufficient credit records, the small- or micro-loan oriented P2P lending companies cannot access their credit records or payment data as freely as traditional financial institutions can (i.e., deposit takers, investors, and insurers). To make it worse, user survey data usually costs a lot of time and money to collect and check, which is unaffordable for P2P lending companies.

[1]http://files.consumerfinance.gov/f/201505_cfpb_data-point-credit-invisibles.pdf.
[2]http://www.bloomberg.com/news/articles/2016-05-26/lufax-ceo-says-ipo-is-probably-a-year-away-amid-market-tumult.

Consequently, online user-generated social data becomes an invaluable alternative to traditional financial data for the P2P lending industry. Thanks to *social-data-based credit scoring*, online P2P lending companies are able to increase approval rates and reduce credit loss with very low cost. Therefore, they are willing to extend credit to grassroots entrepreneurs or household consumers on reasonable terms, which is crucially different from traditional banks or loan sharks.

However, heterogeneous social data such as status updates, social interactions, and the like do not contain explicit credit-related content in most cases [Java et al. 2007; Hua et al. 2015]. Moreover, social data usually contains lots of irrelevant information and even noise about users' credit, thus posing great challenges for traditional credit scoring models. For instance, signals about users' credit risk in the social data are usually quite small and vague. Only after a deep understanding of social data can we make intelligent use of these small signals for personal credit risk evaluation. It is with this motivation that we try to study how to predict this extremely implicit user attribute – credit – by identifying evidence from individuals' digital footprints within their social data. To empirically measure the performance of social-data-based personal credit scoring, we study this problem by employing a carefully prepared Weibo[3] dataset as the testbed. Specifically, the ground-truth credit labels of users in the testbed are already known from our third-party P2P lending partner.

Summarizing results from traditional credit scoring literature, we propose three major principles – Capacity, Character, and Conditions – to guide our solution to the social-data-based credit scoring problem. These three principles are also known as key factors in building traditional consumer credit scoring systems [Rosenberg and Gleit 1994]. By analyzing the most representative words of good and bad credit users, respectively, we will also show that these three principles can be recognized from social data to a large extent. To further bridge the gap between social data and users' credit, we glean the following six credit-related insights: (i) "*economic stability*," (ii) "*experienced employee,*" (iii) "*well educated*," (iv) "*creative poster,*" (v) "*healthy lifestyle*," and (vi) "*prudence and responsibility*" (see Section 3 for details), which are inspired by empirical observations from both good and bad credit users' social data. Our study shows that good credit users tend to possess the aptitudes suggested by these insights, which is very coherent with traditional literature and human intuitions. Based on the proposed principles and insights for social-data-based personal credit scoring, we explore three categories of features extracted from online social data – demographic features, tweet features, and network features – corresponding to the three typical parts of social data. For each feature category, we study the prediction features with feature evaluation metrics like Pearson correlation and χ^2 statistics. Using GBDT [Friedman 2001], we empirically study the relative feature importance within each feature category and the predictive performance of each kind of features. We also analyze the results in detail with the aforementioned principles and insights.

To fully harvest the small signals in social data pertinent to users' credit, we put forward a two-tier ensemble learning framework that integrates all these extremely diverse and weakly credit-correlated social features for credit scoring. In particular, the proposed framework utilizes both stacking and boosting techniques.[4] Experimental results show that, with all effective features as input, our approach achieves a prediction accuracy as high as 58.76% and an AUC value of 0.625 on the large-scale balanced dataset. Although not very high, we should note that even 1% performance improvement in credit risk prediction often means enormous revenues for the financial

[3]http://www.weibo.com, also known as Sina Weibo.

[4]It is worth noting that our previous study [Guo et al. 2016] uses different mining techniques and a different dataset (Cf. Section 7.2 for details).

industry in real life [Blochlinger and Leippold 2006; Einav et al. 2013]. After studying our approach's performance superiority over baselines, we perform case studies to show that our approach produces meaningful results and provides very good interpretability and feasibility. We acknowledge that our work is currently an exploratory study. Several limitations exist before we can fully understand the potentials of social-data-based credit scoring. However, if used properly, our study can provide P2P lending companies tremendous customer value and competitive advantages. Indeed, part of our work has already been deployed in the credit scoring system of our third-party P2P lending partner. We believe that it has good potential of being applied as (i) an indicator to trigger more careful credit checking when official credit records are scant or spotty, and (ii) an auxiliary variable to be incorporated into traditional credit scoring models.

In summary, the main contributions of this work include the following five points:

(1) To the best of our knowledge, we are the first to formally study the problem of inferring user credit labels based on such a broad range of social features. Our investigation is performed on a large-scale dataset from one of the most typical micro-blogging platforms – Weibo (Section 2).
(2) We propose meaningful principles and insights based on traditional literature and empirical observations, which further guide the feature extraction process for tackling the social-data-based credit scoring problem (Section 3).
(3) To efficiently integrate the diverse and weakly credit-correlated social features, we propose a two-tier ensemble learning framework for social-data-based credit scoring. Our approach makes use of both stacking and boosting techniques for ensemble learning (Section 4).
(4) We systematically design three groups of low-level features for boosting as well as high-level features for stacking to capture the latent correlations between social features and user credit. In addition, we evaluate the effectiveness of these features with a variety of importance measurements (Section 5).
(5) Comprehensive experiments are performed to evaluate our social-data-based credit scoring approach armed with both stacking and boosting techniques. The results show that our approach can be 17% more effective than traditional credit scoring methods (Section 6).

The remainder of this article is organized as follows. We present the preliminary analysis and problem formulation of social-data-based credit scoring in Section 2. We present the summarized principles and insights for social-data-based credit scoring in Section 3. We give an overview of our two-tier stacking and boosting enhanced ensemble learning framework in Section 4. After that, we elaborate on the proposed three groups of basic prediction features as well as high-level features in Section 5. Section 6 reports experimental results under different experiment settings. Section 7 details work related to this study. Finally, we draw conclusions and discuss possible future working directions in Section 8.

2. PROBLEM ANALYSIS

In this section, we first present some background and preliminaries of personal credit scoring and give an overview of the Weibo dataset that will serve as the testbed for investigation. After that, we formally introduce the social-data-based credit scoring problem, which is the focus of the remainder of this article.

2.1. Background and Preliminaries

As the name implies, *personal credit scoring* is targeted at evaluating the credit risk of individuals who apply for loans from financial institutions. Different from business loans, personal or consumer loans are often connected with the beneficiary's personal

life. After being granted, individuals often spend the money buying products such as electronics, cars, household items, kids' gear, and appliances. Although the size of these consumer loans is often small, considering the large number of personal financial needs, they are very important for economic growth. For instance, the credit card industry's success originates from the huge demands of consumer finance. It is estimated that "unbanked" consumers will create a $6 trillion consumer debt market globally if increasing numbers of consumers have convenient access to modern financial services through the Internet.[5] In addition, since the founding of the first P2P lending company, Zopa, in February 2005, online P2P lending has proved revolutionary in the financial industry, with transactions taking place totally online. For example, the market scale of online P2P lending in China alone was 103.6 billion RMB in 2014, estimated to increase to 2 trillion RMB by 2024.[6]

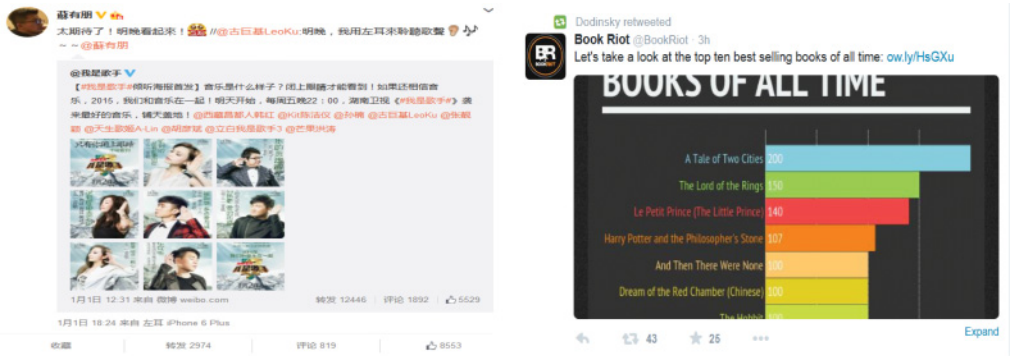
Traditional consumer credit scoring methods have been based essentially on independent variables from the following categories: (i) transaction data characteristics, (ii) consumers' historical financial and credit records, (iii) consumers' demographic information, (iv) product characteristics, and (v) consumers' financial attitudes [Adams et al. 2007; Agarwal et al. 2009; Vissing-Jorgensen 2011; Karlan and Zinman 2009]. However, online P2P lending companies like LendingClub,[7] Kabbage,[8] and Renrendai,[9] cannot access traditional financial data freely, while traditional personal credit scoring methods cannot keep pace with the fast development of today's consumer financial industry. Recently, social data has begun to be leveraged to alleviate the data shortage problem of P2P lending companies. As mentioned in Section 1, applying social data for personal credit scoring becomes increasingly important for PSP lenders. Here, to glimpse the high heterogeneity and complexity of social data, we elaborate on the following three categories of social data typically available on social media platforms:

1) **User Demographic Attributes.** Typical ones include gender, age, education, occupation, hometown location, and so on. This information is usually self-reported by users on the social media.
2) **User-Generated Content.** Included here are the unstructured data generated by users such as texts (micro-blogs, comments etc.), images, videos, and so forth. Mining is primarily targeted at sentiment polarities, posting time, usage of hashtags, language styles, N-gram features, and the like.
3) **User Social Network.** A user's social network data include relationships of friends,[10] followers and followees, as well as ego-network structures.

In particular, we will explore this problem under the setting of micro-blogging platforms, which are among today's most popular social networking platforms and cover all the above-mentioned categories of social data. Without loss of generality, we use the dataset from Weibo, the most popular micro-blogging and social networking platform in China, to study the personal credit scoring problem. The ground-truth credit labels of users are derived from their corresponding financial transaction records, such that a person is of a "good credit" class, if and only if he or she has never previously defaulted on any transactions, which is a common practice in financial literature.

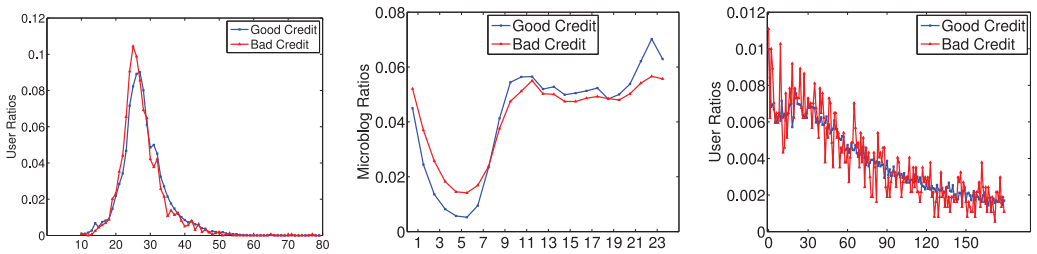
As illustrated in Figure 1, unlike Tweeter micro-blogs, Weibo micro-blogs include plain texts, mentions, hashtags, embedded URLs, and attached images as well as

[5]http://www.wired.com/2013/01/techs-hot-new-market-the-poor/2/.
[6]http://www.boaoreview.com/news/2015/0416/674.html.
[7]https://www.lendingclub.com.
[8]https://www.kabbage.com.
[9]https://www.renrendai.com.
[10]Here we call bi-directional relationships as friends.



(a) A Weibo micro-blog (b) A Twitter micro-blog

Fig. 1. Screenshots showing micro-blog examples from Weibo and Twitter.



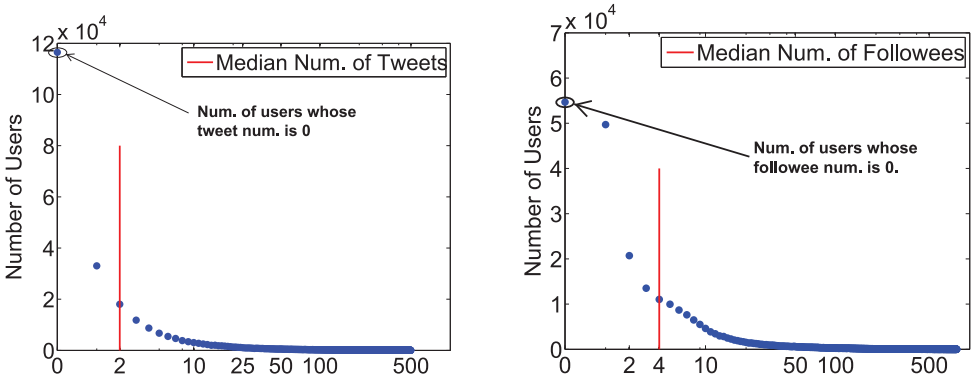
(a) Age (b) Posting time(Hours) (c) Follower No.

Fig. 2. Distribution comparison between good and bad credit users with respect to (a) Age, (b) Posting time of tweets, and (c) Number of followers.

emoticons, retweet chains, and videos. In addition, the Weibo profile page is much more comprehensive than that of Twitter. In genearal, the Weibo profile pages contain users' demographic information such as gender, age, location, verification, marital status, education background, and working experience. In Figure 2, we present some preliminary comparisons between good and bad credit users in the Weibo dataset. The three features are from the three categories of social data mentioned earlier. It can be observed that although the curves of both good and bad credit users show very similar variation trends, the distributional differences between them are not negligible. In Figure 3, we show the distribution of users with respect to number of tweets and number of followees. Both number of tweets and followees per user follow the power-law distribution, which is in accordance with common statistical properties of the social network. It is obvious from Figure 3(a) that more than half of the users have less than 2 tweets. Only 116,478 users (54.4%), composed of 109,455 good credit users and 7,023 bad credit ones, have more than one tweet. It is essential to have enough tweets for each user before feature extraction. After empirical sensitivity studies, we set the minimum number of tweets per user to 21. In total, we are left with 28,830 (95.0%) good credit users and 1,507 (5.0%) bad credit users. In the following studies, we will focus on this subset of 30,337 users to evaluate our social-data-based credit scoring approach.

2.2. Problem Formulation

Based on the preceding analysis, we formulate the *social-data-based personal credit scoring* problem as follows:



(a) No. of Tweets (b) No. of Followees

Fig. 3. Distribution of users with respect to number of tweets and number of followees per user.

Given a set of social data associated with an individual $u_k \in \mathcal{U} = \{u_i\}_{i=1}^n$, classify the person's credit label y_k into one of the two credit risk classes $Y = \{0, 1\}$ (i.e., "good credit" and "bad credit"). Types of social data include profiles, tweets, and social networks. When P2P lending companies evaluate a new customer's creditworthiness, if only social data is considered as input, they are facing the same situation as social-data-based personal credit scoring. Note that it is both impractical and inconvenient to directly assign credit scores to users in the training data. To better utilize statistical tools, most previous credit scoring literature formulates it as a binary classification problem [Thomas et al. 2002]. We follow this convention in our study. The credit scores can usually be obtained by post-processing probability estimates output by the corresponding binary classifiers.

3. INSIGHTS AND PRINCIPLES FOR SOCIAL-DATA-BASED CREDIT SCORING

In this section, we present the three principles that will serve as guidelines for our exploration of social-data-based credit scoring. Moreover, we give a detailed and formal explanation of the six credit prediction-oriented insights that arise from empirical data investigations.

3.1. Credit Scoring Principles and Insights

Traditionally, credit scoring systems make use of data relating to the 5Cs of credit: *C*apacity, *C*haracter, *C*apital and *C*ollateral, and *C*onditions [Rosenberg and Gleit 1994]. "Capacity" refers to users' financial capacity to repay the credit; "character" refers to users' willingness to repay credit; "capital and collateral" refer to the possessions or equities from which payment might be made; and "conditions" refers to the general environment or special conditions applying to the borrowers or the credit types. The 5Cs are general rules for solving the credit scoring problems in traditional literature. In our social-data-based setting, although the private information pertaining to "capital and collateral" cannot be discovered from publicly accessible social data, we will show that the remaining 3Cs *(C*apacity, *C*haracter, and *C*onditions) can be evidenced from social data to a large extent. We summarize them as the three principles for social-data-based credit scoring as follows:

PRINCIPLE 1 (CAPACITY). *Good credit users are more willing to share moments about their personal lives on the social platforms. Some of these moments suggest that they*

Table I. Our Proposed Principles for Social-data-based Credit Scoring and Their Corresponding Representative Insights

Principles for Social-data-based Credit Scoring	Corresponding Representative Insights
Capacity	"Economic Stability" "Experienced Employee" "Well-Educated"
Character	"Creative Poster" "Prudence and Responsibility"
Conditions	"Well-Educated" "Healthy Lifestyle"

are capable of paying back the credit debt in time. Their economic capacity is usually very stable for meeting their payments.

PRINCIPLE 2 (CHARACTER). *Good credit users are more likely to exhibit characteristics indicating that they are content contributors rather than consumers on social media. They also have the characteristics of being prudent and responsible, reflected from their writing styles and content qualities.*

PRINCIPLE 3 (CONDITIONS). *Good credit users maintain good mental and physical conditions, ensuring that no external misfortunes like unemployment or ill-health happen to them in the future. Good health improves one's ability to repay the credit to at least some degree.*

To harness social data for credit scoring, practitioners in the P2P lending industry believe that when extending credit, a person's social standing, online reputation, and/or professional connections are key factors that should be considered,[11] reflecting the critical "C" of *C*apacity. From our practical point of view, the *C*haracter and *C*onditions of an individual can also be inferred from the online social data to some degree using carefully designed social features. As a whole, these three principles are not only supported by traditional literature but also evidenced by observations from our testbed. In total, we glean six insights related to credit scoring through empirical observations of the Weibo dataset: "economic stability," "experienced employee," "well-educated," "creative poster," "healthy lifestyle," and "prudence and responsibility."

All the proposed insights can somehow be explained by these three principles. In Table I, we summarize the relations between the proposed three principles and six insights. As mentioned before, these six insights are based on empirical observations from the Weibo dataset, while the proposed principles are obtained from traditional credit scoring literature. In the following, we elaborate on these social data-driven insights and show that these insights are supported by the 3Cs of *C*apacity, *C*haracter, and *C*onditions.

—Insight "Economic Stability." Intuitively, good credit users should have a stable income every month, which is the first guarantee for repaying the loan on time. For example, they should have stable work prospects for the future. This insight is directly related to the user's ability to pay back their credit (i.e., "capacity')'. On social media, some people constantly tweet about affairs they encounter at work or mention their professions in tweets. These kinds of tweets contain lots of interesting hints about their working "conditions," indicating where they are employed and what their jobs are. After empirical studies, we observe that users with more tweets about working are much more likely to be good credit. We hypothesize that these tweets indirectly reflect people's economic stability.

[11]http://knowledge.wharton.upenn.edu/article/the-social-credit-score-separating-the-data-from-the-noise/.

—Insight "Experienced Employee." Good credit users tend to work at certain jobs for a long time, and be expert in certain areas. Normally, individuals employed by a famous company will have a much more stable monthly income than those making money from small or unknown enterprises. As a result, the longer the user is employed by certain companies or organizations, the better creditworthiness he or she has. Furthermore, experienced employees are very proficient at their work, usually hold higher positions than ordinary workers, and therefore have better "capacity" for paying the loan on time. On social media, if we can infer one is a senior worker at a given company, we can assume that one's credit risk is very low.

—Insight "Well-Educated." As a rule of thumb in credit scoring, a good education or a high academic degree enhance one's likelihood to keep up with payments. It is widely accepted in the real world that more education means better reputation, better social standing, and even better overall performance in life and work. For example, well-educated individuals are more likely to be prepared for accidents in life, and well-educated employees have more chances and potentials to be promoted during their careers. Usually, if we know a given user is well-educated, we can usually predict that her credit default risk is very low. This insight is simultaneously correlated with "character," "capacity," and "conditions." Note that we can also indirectly conclude one's education level from language style and tweet topics on social media, which are accurate and convincing in inferring one's intellectual development.

—Insight "Creative Poster." Good credit users tend to spend more time posting and like to share personal affairs with their friends, rather than retweet or tweet about news, reviews, old sayings or quotes on social media. After empirical comparisons, we find that good credit users tweet about their everyday lives much more often than do bad credit ones. They are somewhat creative posters, recording their thoughts and feelings frequently on social platforms. This phenomenon indicates that good credit users are having a very positive attitude toward life and work. On the contrary, we find that a number of bad credit users only treat social media as another channel to learn about news, express opinions, or post comments. Although some of them are also active participants in online social activities, they mostly act like retweeters and commenters. In short, we can infer that (i) good and bad credit credit users have different "characters" and attitudes in terms of social engagement, and (ii) good credit users are skilled at the sophisticated functions of social platforms and prefer to create original posts rather than retweet or comment.

—Insight "Healthy Lifestyle." The fact that tweets can reveal one's health status has been well-recognized in the past few years [Paul and Dredze 2011]. Figure 2(b) shows the comparison of posting time distribution between good and bad credit users. We can see that bad credit users tweet more during the early hours, while good credit users tend to tweet during daylight or evening hours. The posting time distribution is a strong indicator of online users' activity intensity over days and nights. It seems that bad credit users are more likely to stay up very late. In addition, we also observe that a percentage of bad credit users often talk about suffering from ailments such as insomnia or flu within their social updates, further implying their bad health status. Broadly speaking, good physical and mental conditions are the critical "conditions" of steady performance in life and work. If a severe illness happens, bad health will immediately deteriorate the borrower's overall "conditions," and thereafter the borrower's credit risk increases.

—Insight "Prudence and Responsibility." In addition to the capacity for repaying the loan, good credit users tend to possess the personality of being prudent and responsible. This kind of user is usually more concerned about the rules of modern society and therefore performs well in keeping promises and maintaining creditworthiness. We propose this insight from the intuition that some bad credit users may default

Table II. Insights Inferred from Empirical Observations and Corresponding Representative Social Data Features

Insights Synthesised from Empirical Observations	Corresponding Representative Features Extracted from Social Data
Economic Stability	Age Occupation types
Experienced Employee	Number of years since the user starts his or her career Number of companies where the user has worked
Well Educated	Education level Sentiment vocabulary (e.g., vulgar language)
Creative Poster	Usage of Emoticons Average length of retweet chains
Healthy Lifestyle	Fraction of tweets published at each hour during the day Sentiment polarity distribution
Prudence and Responsibility	Number of duplicate tweets Aggregated features of one-hop neighbors' degree features

simply because they are careless and forget to pay back their debt in time. In this case, these bad credit users are capable of repayments but are unaware of the default risk. We think that this personality or "character" can be reflected in the use of ill-spelled words in their postings, frequent retweets with no comments, and the like. In addition, we also observe that bad credit users often participate in activities of App marketing or other kinds of product commercial campaigns that offer prizes or lucky draws as incentives. Frequent involvement in these activities suggests that they are fond of small bonuses and try their luck on the Internet. It is probably true that this personal "character" degrades their credit worthiness.

In Table II, we list the six data-driven insights with their corresponding representative features, which will be detailed in Section 5. We can see that the features in the right column clearly support insights in the corresponding left column. In sum, the proposed principles and insights can serve as good guidelines for feature extraction for social-data-based credit scoring. In the following, we show some empirical evidence that supports these principles and insights.

3.2. Empirical Evidence of Principles and Insights

In this subsection, we present some exploratory analyses on the Weibo testbed to verify the principles and insights just mentioned. Empirical evaluations in Section 5 will further show that features inspired by these proposed principles and insights are effective in discerning good credit users from bad ones.

The textual content generated by users is very important and useful in understanding the differences between good and bad credit users. We propose analyzing both good and bad credit users using their most predictive and representative words. These words can be identified using various classification models with unigram features as input. To be specific, these predictive words correspond to the unigram features whose weights are maximally negative or maximally positive in the learned model. Here, we adopt Naive Bayes and Logistic Regression as the classifiers since they are suitable for high-dimensional unigram features. For the sake of fairness between good and bad credit labels, the classification model is learned from a balanced dataset where the numbers of good and bad credit users are equal.

Table III shows the most predictive words learned by the multinomial Naive Bayes model, which has been proved to be very competitive in Twitter user classification [Hong and Davison 2010]. Based on words like "laughing," "sleep," and "car," we can see that good credit users like to tweet about what they find and how they feel in their everyday lives, while the remaining words strongly reflect users' thoughts and feelings about what they have finished in the past and plan in the future. These words suggest that

Table III. Most Predictive Words for Both Good and Bad Credit Users' Output from Naive Bayes Model Using the Words of Users as Features

Classes	Most Predictive Words
Good Credit	laughing, sleep, tomorrow, dear, finally, should, having the guts, hour, car, slightly, Shanghai, recently, prepare, really, afterwards, at last, future, send etc.
Bad Credit	luck, money, sponsor, prize, win, address, participate, come on, from, wish, recommend, obtain, mood, blogs, free, game, lucky, more, iphone etc.

Table IV. Most Predictive Words for Both Good and Bad Credit Users' Output from Logistic Regression Model Using the TF-IDF Features

Classes	Most Predictive Words
Good Credit	photograph, teacher, eat, husband, reply, street, miui, buddy, log in, have a look, training, XiaoMi, baby, works, mv, camera, Japan, Beijing etc.
Bad Credit	address, shake, hello, contact, short message, high way, poker, busy, app, game, expert, rewards, hyperlinks, world cup, forever, fun etc.

they are optimistic about themselves, lead a healthy life both physically and mentally, and are busy working or planning for the future. On the contrary, bad credit users tend to participate in online advertising activities, which often provide "lucky" draws like an "iphone" as incentives for spreading the influence of goods or services, promoting adoption of Apps, and the like. Another personality trait of bad credit users we can infer from these words is that they are fond of playing online games. These personality traits indicate that bad credit users are not economically well-off and are interested in gaining small bonuses without effort rather than taking systematic actions and investing time and energy in improving their skills for making money. Observations from these predictive words consistently support our intuitions on the principles of CAPABILITY, CHARACTER, and CONDITIONS.

With TF-IDF features as input, we can gain more insights into people's characteristics from less frequent but informative words in texts. Table IV lists the most predictive words for good and bad credit users learned from the L_1-regularized logistic regression model, where the input unigram features are weighted using the TF-IDF strategy. For good credit users, we can find that (i) good credit users often mention taking "photographs" or filming with "cameras" using fashionable smartphones like "XiaoMi"; (ii)good credit users tend to participate in "training" programs and may attempt to gain skills from "teachers"; and (iii) good credit users tend to travel between metropolises like "Beijing" and "Shanghai" and even go abroad to countries like "Japan." These characteristics demonstrate that good credit users are economically well-off since they have access to high-quality products, training, and travel. They are also creative posters, sharing their everyday lives on social media. With regard to bad credit users, we can observe from Table IV that (i) they like to have "fun" playing "games" like online "poker" that are prevalent on the Internet; and (ii) they use words indicating that they are active in commercial propaganda to win "rewards" and share relevant "hyperlinks." These observations indicate that bad credit users might spend lots of effort on online games. In a word, we can confirm that good credit and bad credit users have distinguishable characteristics revealed by the social data, which verifies the credit scoring principles of CAPABILITY, CHARACTER, and CONDITIONS to some extent.

4. FRAMEWORK OVERVIEW

In this section, we present the details of our feature-based two-tier ensemble learning framework, which is purposed to tackle the personal credit scoring problem using heterogeneous social data as input. For Tier-1 classifiers, we adopt classification algorithms including Naive Bayes, Logistic Regression, and SVM. To implement stacking, a state-of-the-art ensemble learning method [Rokach 2010], Tier-1 classifiers first build models based on part of the training data and then make predictions on

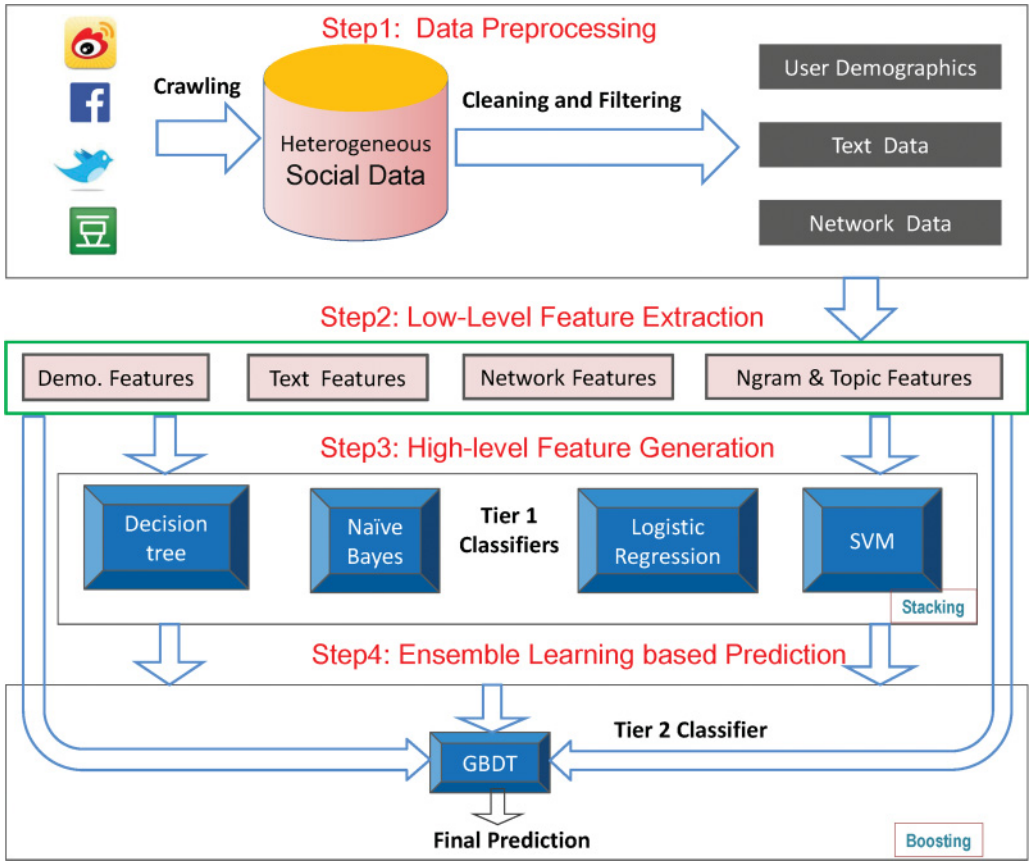


Fig. 4. Illustration of our two-tier ensemble learning framework.

the remaining training data to generate high-level features. For Tier-2 classifier, we adopt the well-known ensemble learning algorithm called Gradient Boosting Decision Tree (GBDT) [Friedman 2001], which consists of an ensemble of fitted regression trees. As illustrated in Figure 4, our framework's pipeline is mainly composed of four steps:

1) Data Preprocessing. For data preprocessing, we remove duplicate tweet records, tokenize text data, and filter out users with too few tweets. To be specific, we set the minimum threshold to 21 tweets per user in experimental evaluations. The dataset of users whose number of tweets is lower than 21 can be exploited for training Tier-1 classifiers, which further generates high-level features for users with more than 21 tweets. After data preprocessing, we obtain user demographics, tweet data, and network data for each user from the social data.

2) Low-level Feature Extraction. In this step, we extract features from three aspects of the data available in our dataset: demographic features, tweet features, and network features as well as ngram and topic features. As illustrated in Figure 4, low-level features are used as the input of Tier-1 classifiers for stacking and are also input for the Tier-2 classifier. How to extract these features and detailed analyses on these features are elaborated in Section 5.

3) High-level Feature Generation. For some extremely high-dimensional features like ngrams, it is time-consuming and inefficient to directly feed them to GBDT. To that end, we propose to first combine high-dimensional features into single high-level

feature with the stacking technique, which trains Tier-1 classifiers with a sampled dataset. The predictions of learned Tier-1 classifiers on the rest of training data, such as predicted labels or estimated probabilities, are used as the high-level features. Apart from the formally defined ngram features, we also include topic features as well as low-level features from Step 2 for stacking. The combination algorithms for low-level features correspond to the Tier-1 classifiers illustrated in Figure 4. Similar to other ensemble methods, the stacking technique also combines the advantages of different Tier-1 classifiers to some extent for the final prediction.

4) Ensemble Learning-based Prediction. To integrate different types of features into a unified credit scoring model, we choose the ensemble learning method GBDT as the final prediction classifier for its outstanding speed, stability, and accuracy in performance. In our ensemble learning framework, GBDT is treated as the Tier-2 classifier. GBDT requires no data normalization before training and handles missing values and nonlinear relationships for high-dimensional data naturally. As mentioned in Step 3, we use stacking to handle the diversity and heterogeneity of social features, which is also a type of ensemble learning method. As illustrated in the lower part of Figure 4, the input of the Tier-2 classifier consists of both low- and high-level features.

GBDT is much more competitive and practical than other ensemble methods like Random Forest [Breiman 2001] because of its gradient boosting methodology, which optimizes certain loss functions in an iterative gradient-descent fashion. To be specific, the loss function of our framework's GBDT model is logistic loss with L_1-regularization as denoted in Equation (1):

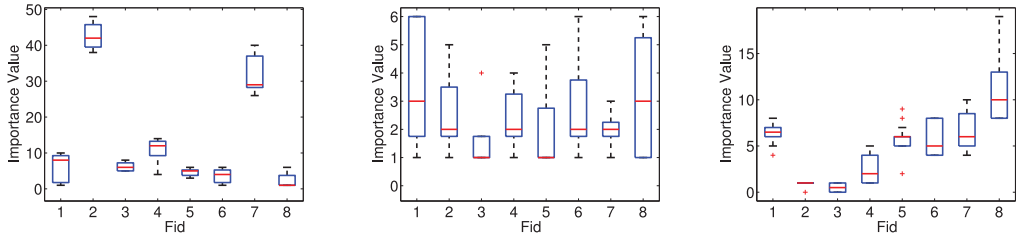
$$\min_{w} \alpha \|\boldsymbol{w}\|_1 + \sum_{i=1}^{n} log(1 + e^{-y_i \boldsymbol{w}^T \boldsymbol{x}_i}), \tag{1}$$

where y_i represents the credit label of user u_i, x_i represents the input features of u_i, and $\|\boldsymbol{w}\|_1$ represents the 1-norm of parameter vector w. The L_1-regularization term $\|\boldsymbol{w}\|_1$ in the loss function can shrink many of the regression coefficients of $\boldsymbol{w}$ to zero and therefore perform feature selection implicitly during model construction. Our experimental results show that L_1-regularized logistic loss is the best loss function to use in GBDT for our social-data-based credit scoring task. After tuning with cross-validations, we set the parameter $\alpha = 1$.

5. PREDICTION FEATURES

In this section, we present our proposed prediction features and describe how to extract them. The set of all low-level features can be divided into demographic features, tweet features, and network features, which together compose the basic input of our credit scoring model, while high-level features comprise features based on predictions of Tier-1 classifiers including Naive Bayes, Logistic Regression, SVM, and Decision Tree. These classifiers are trained with appropriate features like unigrams and topic distributions, as well as low-level features. Table V shows a summary of the social features we use for credit prediction.

Moreover, we will analyze and study the effectiveness of these features with various methods. First, we show the comparison of relative feature importance for each feature, corresponding to the number of times the feature is chosen as the node for splitting when building the GBDT model. The feature importance values output by GBDT are often more consistent with the final results. Second, to demonstrate the effectiveness of these features for credit prediction, we compare our proposed features using feature evaluation criteria such as Pearson correlation and χ^2 statistics. Last, we perform 10-fold cross-validation on a sampled dataset with 1,500 good and 1,500 bad credit users, where GBDT is used as the classifier, to show the effectiveness of these features. In



(a) Demographic features (b) Tweet features (c) Network features

Fig. 5. Feature importance comparison among groups of features listed in Tables VI, VII, and VIII using box plots.

this way, we present performance comparisons among the features of each group and report the comparison results in terms of Precision, Recall, and F1-Score. To evaluate each feature's predictive power fairly, we show the prediction results in terms of good and bad credit users, respectively. As shown in the following figures, the performance is not very high using only one kind of features as input. We will also show that these prediction features are in accordance with the proposed insights and principles in Section 3.

It is worth noting that some features may include a set of related definitions or lots of dimensions. For the sake of brevity, we only present the results of the most representative definition or dimension for the corresponding features. This notation usage applies to all the following feature analysis. As shown in Table I, the proposed insights and principles are correlated with each other, so we only focus on analyzing and discussing the relationships between insights and features in the section.

5.1. Demographic Features

Our demographic features are extracted essentially from Weibo user profile pages, which provide input fields about users' personal profiles including screen name, gender, age, location, education background, working experience, interest tags, and registration time. Compared to Twitter, Weibo offers a much more comprehensive description of users' personal profiles. As a result, we can obtain much more precise and fine-grained demographical features than previous studies.

5.1.1. Feature Definition. We define the representative demographic features based on the data fields on Weibo as follows:

Screen name: The screen name is a unique identification string for users to represent themselves on social media. A lot of character patterns and semantic information can be extracted from screen names. We define screen name features including length of the screen name, number and proportion of alphabetic characters in the screen name, number and proportion of numerical characters in the screen name, and number and proportion of symbol characters in the screen name.

Gender and Age: Normally, users enter their gender information in the profile page. Because some do not disclose gender online, their gender fields are empty. We define the gender feature based on self-reported information, which takes trinary values of {male, female, unknown}. The age data field has much in common with that of gender. We define the age feature as how old the user is.

Verification: Verification is a community-fostering function provided by the Weibo platform to verify the identity of users, and this is useful to promote visibility and attract followers. If one wants to be verified, one has to submit materials about personal identity or career to Weibo. After Weibo's approval, a verified title will be placed under

Table V. Summary of Social-data-based Features for Personal Credit Prediction

Feature Group	Feature Descriptions
Demographic features	Length of the screen name
	Number and proportion of alphabetic characters in the screen name
	Number and proportion of numerical characters in the screen name
	Number and proportion of symbol characters in the screen name
	Gender and Age of the user
	Whether the user's identity is verified by Weibo or not
	Education level of the user
	Provinces where the user lives
	Number of companies where the user has worked
	Number of years since the user starts his or her career
	Whether the company the user works in is renowned or not
	Number of years and months since the user joined Weibo
	Active level of the user
Tweet features	Number and fraction of retweets of a user's tweets
	Number and fraction of retweets with no comments
	Average depth of retweet chains
	Maximum depth of retweet chains
	Depth deviation of retweet chains
	Number of emoticons/mentions in users' tweets
	Standard deviation of number of emoticons/mentions in a user's tweets
	Average number of emoticons per tweet
	Fraction of tweets that contain emoticons/mentions
	Fraction of tweets at each of 24 hours of a day
	Number and fraction of tweets whose sentiment polarities are positive/negative/neutral respectively
	Deviation of the sentiment polarity values among users' tweets
	Number of positive/negative sentiment word occurrences in users' tweets
	Fraction of positive/negative sentiment words in users' tweets
Network features	#followers, number of followers
	#friends, number of friends
	#friends/#followers, fraction of followers that are also followees
	#friends/#followees, fraction of followees that are also followers
	#followers/#followees, fraction between number of followers and followees
	Aggregated values of a user's one-hop neighbors' network features
	Betweenness Centrality
	PageRank values
High-Level features	Features derived from ngram features using Logistic Regression
	Features derived from ngram features using Naive Bayes
	Features derived from topic distributions using Logistic Regression
	Features derived from topic distributions using Naive Bayes
	Features derived from topic distributions using Decision Tree
	Features derived from demographic features with different classifiers
	Features derived from tweet features with different classifiers
	Features derived from network features with different classifiers

the user's screen name. Because the verification is not uniform in texts, we define verification feature as whether the user is verified or not.

Education: Only about 10% of the users in the Weibo dataset enter education information in their profile forms. Users can input education information at all levels. Because the highest academic institution one has entered is the most important in predicting education level, we define the education feature as the highest degree users have achieved. From institution names, rule-based methods can extract the types of institutions. The possible values that education feature can take include junior high school, senior high school, polytechnic school, and university.

Location: The current residential location of users is another important input field on the Weibo profile page. Due to Weibo's policy, all users fill in the current city and province they are living in when registering. Locations are nominal attributes of users

Table VI. Pearson Correlation and χ^2 Statistics for Demographic Features

Fid	Feature Name	Pearson Correlation	χ^2 Statistics
1	Gender	4.45×10^{-2}	14.27*
2	Age	1.92×10^{-2}	16.28*
3	Verification	5.128×10^{-2}	**17.02***
4	Education	4.18×10^{-3}	0
5	Location	4.81×10^{-2}	16.68*
6	Occupation	2.244×10^{-2}	0.137~
7	Registration time	$\mathbf{6.944} \times 10^{-2}$	39.44*
8	Active Level	4.770×10^{-2}	31.77*

*Passes significance test at the confidence level of 95%.

and have no ordering among different values. As a result, we define separate binary values derived from users' locations as features. More specifically, experiments show that the predictive performance of city features is slightly lower than that of province features, which are much denser. We define the location features from the provinces users entered.

Occupation: Similar to education, lots of users do not reveal their occupation information in detail. This kind of feature suffers from data sparsity too. We define the following three simple features from the working experience users have provided: number of companies worked in; number of years since the user started her career; whether the company the user works in is renowned or offers a high salary. In most cases, company names are informal and vary due to users' preferences, so we do not directly use company names as features.

Registration time: This field is automatically generated by Weibo, so every user has an exact registration time on the profile page. Since Weibo started operations in 2009, the registration time cannot be older than 2009-01-01. We define features including the number of years and months since the user joined Weibo.

Active level: We directly adopt Weibo user activity statistics, including the current active level of the user, the accumulated active days of the user, additional active days to upgrade active level, and time that has elapsed since the user's last login.

5.1.2. Analysis and Discussion. In Table VI, we show the statistical comparison between different demographic features with respect to Pearson correlation and χ^2 statistics. Features from "Screen name" are not included for comparison because there are almost no differences between good and bad credit users on these features, despite the fact that a screen name is very personalized for each user. For the remaining features, the low Pearson correlation values demonstrate that there only exist weak linear dependencies between demographic features and users' credit labels. However, as shown in Table VI, most demographic features pass the significance test by χ^2 statistics. It is rather counterintuitive that features like education and occupation fail to pass the significance test. The reason could most probably be attributable to the severe data sparsity issues mentioned earlier. For example, the missing value ratios for features "Age," "Occupation," and "Education" are 69.78%, 93.14%, and 89.66%, respectively.

In Figure 5(a), we can see that (i) feature importance distribution is not always consistent with statistical analysis in Table VI; and (ii) the features "Age" and "Registration time," with overall feature importance values as high as 40 and 30, respectively, exhibit much greater predictive power than the remaining demographic features. In Figure 6, we present the primary prediction results of these features. Because features with lots of missing values could not perform very well, we only present the results of "Verification," "Location," "Registration time," and "Active level" for fairness. Figures in the first row show the Precision, Recall, and F1-Score values with respect to good credit

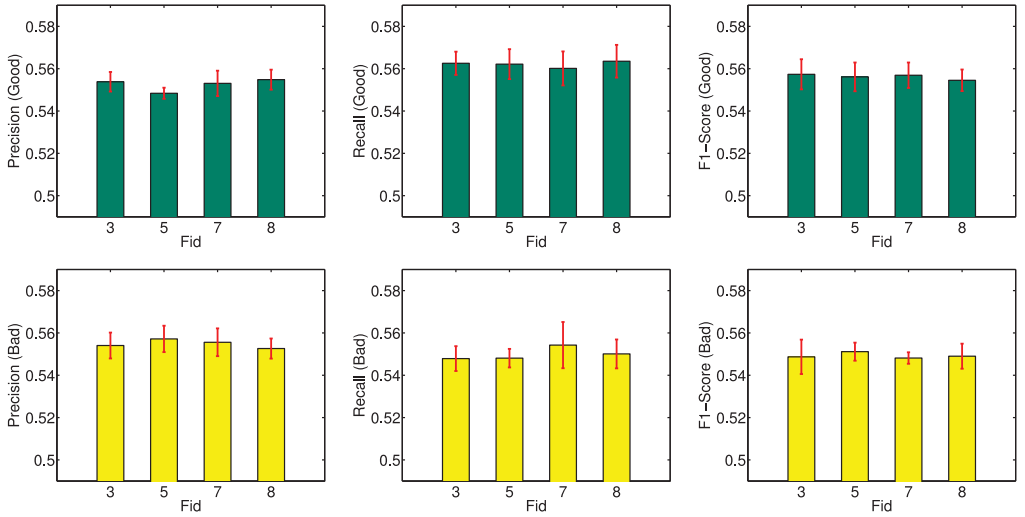


Fig. 6. Prediction results for both good and bad credit users with different features in Table VI as the input.

users, and figures in the second row with respect to bad credit users. Surprisingly, we can see that the overall performance of these features is comparable, ranging from 0.54 to 0.56. Although the standard deviations are not negligible in Figure 6, the mean values are well above 0.5. All these phenomena demonstrate that these features are discriminative in personal credit scoring.

Among the demographic features, "Gender," "Age," "Education," "Occupation," and "Active Level" confirm their importance with their strong predictive power in the preceding effectiveness analysis. For instance, the feature "Age," with a missing ratio of 69.78%, still has the highest feature importance value in Figure 5(a). The reason may be that "Age" and "Occupation" reflect users' working experiences, and senior/older workers are more likely to be experienced employees. "Location," whose χ^2 statistics is the highest in Table VI, reflects the economic development imbalance of different geographic regions. In particular, our primary data analysis also reveals that certain regions do have a higher frequency of credit fraud. "Registration time" could tell if a user is an early adopter of new technology like Weibo. Mostly, the early adopters belong to the intellectuals of society. Although some may still be students or in a low-income population, we can at least predict that early adopters of new technologies are more likely to perform well in life and work.

To sum up, these demographic features are well in accord with insights like "Economic Stability," "Experienced Employee," and "Well-Educated." The feature "Active level" is a good indicator of users' engagement with the Weibo platform and is correlated with insight "Creative Poster." It is also worth noting that a number of other demographic features like number of badges awarded by Weibo and number of status updates, are available on the profile page. But after performing feature evaluations, they show no discriminative power in credit prediction. Similar to Twitter, Weibo also offers a short bio field for brief self-introduction, but we find scarce informative profile information in it since the Weibo profile page already contains adequate demographic input fields.

5.2. Tweet Features

Aside from demographic data, tweet data are also available on the social platforms. This kind of data, although not directly related to users' credit attributes, reflects

users' personal preferences and habits very well, which can also predict their financial behavior to a large extent.

5.2.1. Feature Definition. Under the setting of micro-blogging platforms, we consider the following features related to social users' tweets.

Duplicative behavior: There are a few cases in which users post tweets that are almost the same as others' tweets (e.g., those they have previously published or simply copied from others). Duplicative behavior reflects how creative the user is in generating content and, to some extent, how careful she is in maintaining a personal image among friends. We use the SimHash algorithm proposed in Charikar [2002] to identify duplicative tweets and extract duplicative behavior features, including the number and fraction of duplicated tweets for each user.

Retweet behavior: Retweet behavior habits reveal a lot about a person's personality traits on social platforms. We extract as features the number and fraction of retweets for each user. We also extract retweet chain features, including the average and maximum length of retweet chains and the standard deviations of retweet chain length. If the retweet has no comment, we call this a *plain retweet*. The above- mentioned retweet behavior features can be further divided into two cases depending on whether these retweets are plain retweets or not.

Usage of emoticon and mention: The usage of emoticons and mentions is prevalent on Weibo. We define features related to emoticon usage as the number of times emoticons are used in the user's tweets, the average number of emoticons per tweet, and the fraction of tweets that contain emoticons. We also extract three similar features for the use of mentions.

Posting time: The posting time of tweets tells a lot about the author's characteristics, and these turn out to be very informative features between users of different credit classes. We define the fraction of tweets published during each hour of the day as the posting time feature, which can further be divided into two cases according to whether the tweets are retweets or not.

Sentiment vocabulary: We observe that the sentiment vocabulary used by a user serves as a good indicator of a user's overall sentiment distribution. We first manually label the sentiment words that are unique to the Weibo dataset and are usually emerging words in oral language. With additional public sentiment words, we construct a sentiment vocabulary for feature extraction. Then, we define the following sentiment vocabulary features: the number of occurrences of positive/negative sentiment words and the fraction of positive/negative sentiment words in users' tweets.

Sentiment polarity: Following the methodology proposed in Xiang and Zhou [2014], we train a sentiment classifier on a Weibo dataset with known sentiment labels. We then classify the sentiment polarity of tweets into three classes: "positive," "negative," and "neutral." For each user, we define sentiment polarity features as the fraction of tweets belonging to each of the three sentiment classes and the standard deviation values of sentiment polarity among users' tweets.

5.2.2. Analysis and Discussion. Table VII shows that six of the tweet features, including "Retweet chain," "Plain retweet," "Emoticon usage," "Mention usage," and "Posting time," are statistically significant by χ^2 test. "Posting time" is especially useful in credit prediction since the χ^2 statistics are noticeably high for all 24 features of this kind. In Figure 5(b), we see that the importance values of tweet features are comparable. Although the feature importance of each tweet feature is not high, they collectively demonstrate good predictive performance, as will be shown in our experiments. In Figure 7, we show the direct performance comparison among the different tweet features with respect to Precision, Recall, and F1-Score. Similar to demographic features,

Table VII. Pearson Correlation and χ^2 Statistics for Behavior Features

Fid	Feature Name	Pearson Correlation	χ^2 Statistics
1	Duplicative Behavior	2.740×10^{-2}	2.642˜
2	Retweet Chain	$\mathbf{9.200} \times 10^{-2}$	53.05*
3	Plain Retweet	3.374×10^{-2}	34.61*
4	Emoticon Usage	8.637×10^{-2}	25.68*
5	Mention Usage	6.236×10^{-2}	28.10*
6	Posting Time	5.162×10^{-2}	**61.06***
7	Senti. Vocab.	4.240×10^{-2}	0.380˜
8	Senti. Polarity	9.272×10^{-3}	2.268˜

*Passes significance test at the confidence level of 95%.

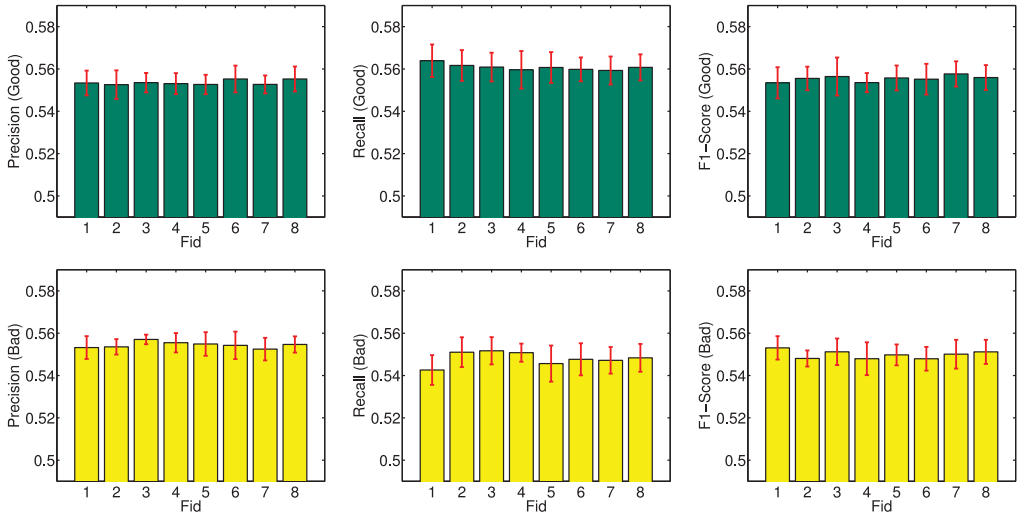


Fig. 7. Prediction results for both good and bad credit users with different features in Table VII as the input.

all tweet features show some predictive power for both credit labels. We can confidently say that these features are predictive of credit worthiness to some extent.

To be specific, "Duplicative behavior" is more prevalent among users who view Weibo as an advertising platform for their products or services. This feature also reflects users' caution about their reputation since repetitive postings may annoy others. Therefore, "Duplicative behavior" features correlate with the insight "Economic Stability" and "Prudence and Responsibility" to some extent. The features of "Retweet chain" reflect a user's social connections with others to some extent, while the "Plain retweet" feature indicates a user's creative level when propagating online events and topics. The retweet behavior is coherent with the insight "Creative Poster."

The "Posting time" features reveal interesting behavioral differences between good and bad credit users – bad credit users tend to post status updates late at night (i.e., from 00:00 AM to 7:00 AM), while good credit users' posting time distribution conforms more to normal hours, as shown in Figure 2(b). These late night postings are probably due to a habitual nightlife, insomnia, or an unhealthy lifestyle, which may hurt the creditworthiness of the corresponding user in the long term. This phenomenon can be well explained by the insight "Healthy Lifestyle." "Emoticon usage" features reflect users' proficiency at social media expressions as well as their personal mood variations, while "Mention usage" features indicate their affection for the "mention" function on social media. Both kinds of tweet features reflect the insight "Creative

Poster." The sentiment features capture users' overall sentiment polarity and mental stability, corresponding to the insight "Prudence and Responsibility." For example, a prudent and responsible person seldom uses vulgar language to express bad sentiments like anger on social media.

In this study, we also tested more than 20 other kinds of tweet features, such as badge-related tweets, usage of punctuation and symbols, and the like, but results show that these features are not discriminative between good and bad credit users. We omit the discussion of these features for brevity.

5.3. Network Features

Compared to traditional financial data, online social network data directly reflect users' social status. On Weibo, the profile page provides basic statistics of users' social connections, including number of followers, number of followees, and number of friends. The social network structure is another unique information source on social platforms. Although we cannot access the whole social network of Weibo users, we can obtain a given user's one-hop network structure after her authorization. We call this one-hop network structure the *ego-network*of a given user. We denote the total set of ego-networks of users by $\mathcal{G} = \{G_i = (V_i, E_i)\}_{i=1}^{n}$, where n is the number of users.

5.3.1. Feature Definitions. Degree features: The degree of users in social network is the basic measurement of one's popularity and gregariousness. We define degree features based on the basic statistics of a given user u_i as follows: # of followers (i.e., the number of followers); # of followees (i.e., the number of followees); # of friends (i.e., the number of friends); # of friends/# of followers (i.e., the fraction of followers who are also followees); # of friends/# of followees (i.e., the fraction of followees who are also followers); and # of followers/# of followees (i.e., the ratio between the number of followers and followees).

Ego-network aggregated features: The ego-network G_i of u_i is another important data source for network features. For u_i's one-hop neighbors $V_i \in G_i$, we can also obtain the basic degree features of them from Weibo's open API. The ego-network aggregated features of u_i can be obtained by computing the mean and variance of the corresponding degree feature values of users in V_i. It is worth noting that we can obtain three sets of aggregated ego-network features if we consider u_i's connections, such as followers, followees, and their combination, separately.

Network centrality features: Based on the ego-network structure, we also introduce network centrality features like "PageRank" and "Betweenness Centrality" of users. Both measurements compute the centrality and importance of nodes using more sophisticated algorithms. These features usually present more comprehensive and precise measurements than degree features for users' social status in the social network. PageRank is a variant of the Eigenvector centrality of networks, and Betweenness Centrality value of u can be computed by Equation (2):

$$C_B(u) = \sum_{s \neq u \neq t} \frac{\sigma_{st}(u)}{\sigma_{st}}, \tag{2}$$

where σ_{st} is the number of shortest paths from node s to t and $\sigma_{st}(u)$ is the total number of those paths that pass through node u. As mentioned earlier, we do not have access to the whole social network of Weibo. To compute betweenness and pagerank values, we run the corresponding algorithms on a connected network constructed from the total ego-networks of all 200,000 users. Details about how to obtain ego-networks will be presented in Section 6.1.

5.3.2. Analysis and Discussion. Table VIII and Figure 5(c) present the effectiveness analysis of network features with respect to Pearson correlation, χ^2 statistics, and feature

Table VIII. Pearson Correlation and χ^2 Statistics for Network Features

Fid	Feature Name	Pearson Correlation	χ^2 Statistics
1	Degree Features	$\mathbf{4.651} \times 10^{-2}$	**23.62***
2	Aggregated Features	2.961×10^{-2}	3.844
3	Centrality Features	2.237×10^{-2}	3.658

*Passes significance test at the confidence level of 95%.

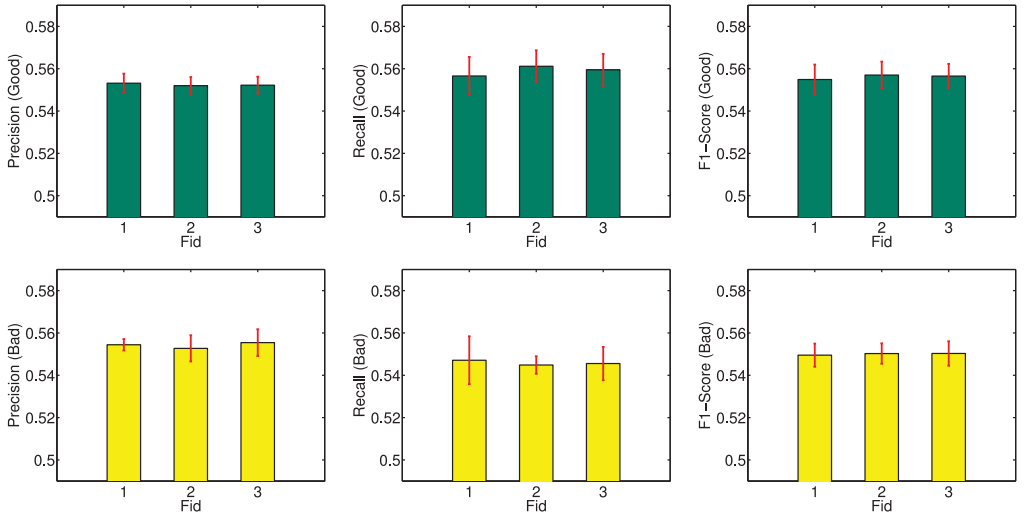


Fig. 8. Prediction results for both good and bad credit users with different features in Table VIII as input.

relative importance. Among the network features, only degree features such as # of followers and # of followees pass significance test, suggesting the weak linear dependence between network features and users' credit labels. But feature importance comparisons in Figure 5(c) show that aggregated features and centrality features are much more important for credit prediction than are degree features. These observations show that all three kinds of network features are informative in personal credit scoring, but they work in different ways. We can also say that ego-network aggregated features circumvent the obstacle of missing a complete network and capture the overall social status of users to a large extent.

In Figure 8, we present the prediction results of these three kinds of network features on the balanced dataset mentioned earlier. To our surprise, network feature performance is very similar to that of tweet features: Although the values of standard deviation are not negligible, different kinds of network features produce comparable results and have a slightly higher performance for good credit users in terms of Recall. In regard to F1-Score, network features also consistently perform better in predicting good credit users. We believe that this phenomenon is a result of the volatile and dynamic nature of the personal credit scoring problem, which is extremely hard to predict for most real-world users.

As noted earlier, features including "# of followers," "# of followees," and "Network centrality" have very good predictive power, which is coherent with our intuition that they measure users' social status. People with high social status are usually expert at something and can attract followers on social media. In addition, people with good reputations usually act with "Prudence and Responsibility" in the online world. We can reckon with high confidence that they have the "Character," "Capacity," and "Conditions" to repay loans on time.

5.4. High-Level Features

In this subsection, we present the set of high-level features that will implement the stacking strategy. The main procedure of stacking includes three steps: (i) sample a subset of the training data,[12] (ii) train a classifier on the sampled dataset, and (iii) apply the classifier to the remaining training data to make predictions that are the corresponding high-level features for our final predictions. It is widely accepted that high-level features help improve algorithms' generalization performance by manipulating input features. In this study, we propose three sets of high-level features derived from ngram features and topic distributions, as well as the low-level features proposed in the previous subsections.

5.4.1. Feature Definition. The details of the proposed high-level features are as follows:

Features derived from unigram features: To exploit user characteristics embedded in textual contents, we derive high-level features from unigram features by applying a multinomial Naive Bayes algorithm to them. We first use a subset of the training data to train a Naive Bayes model, then apply this model to the rest of the training data to predict the credit class probabilities of the remaining users. Finally, we normalize the credit class probability values to "good" or "bad" for each user. In addition, we also derive high-level features from unigram features weighted with the TF-IDF strategy, which favors less frequent words. To use TF-IDF features as input for generating high-level features, we adopt L_1-regularized logistic regression as the classifier. The prediction output of the logistic regression model is already normalized and can directly serve as the high-level features for stacking.

Features derived from topic distributions: Topic distributions exhibited from users' tweets are an important part of user-generated content. We use the Latent Dirichlet Allocation (LDA) [Blei et al. 2003] model to extract users' topic distributions. Disappointedly, for the Weibo dataset in our testbed, directly applying the LDA model to extract topic distributions leads to no discriminative power. We propose to first train an LDA model on another reference corpus composed of more comprehensive and high-quality Weibo tweets. To that end, we first identify 132,846 users, each with more than 10,000 followers and over 1,000 tweets, and then aggregate each user's tweets into one document. To infer LDA topics from the large Weibo corpus, we use the Gibbs sampling algorithm [Griffiths and Steyvers 2004]. For the LDA model, the number of topics $T = 200$, parameter $\alpha = 0.01$, parameter $\beta = 0.25$, and number of iterations $= 500$. After obtaining the LDA model, we can estimate the topic distributions $\Theta_i = \{\theta_{ij}\}_{j=1}^{T}$ of user U_i in our testbed by:

$$\theta_{ij} = \frac{\alpha + \sum_{i \in U_i^{new}} n_{ij}}{\sum_{j=1}^{T} (\alpha + \sum_{i \in U_i^{new}} n_{ij})}, \tag{3}$$

where U_i^{new} denotes the set of words in new user U_i's tweets and n_{ij} denotes the number of times topic j is assigned to word i in the previously learned LDA model. After we infer the topic distributions Θ_i for each user u_i, we can apply classifiers such as SVM and Logistic Regression to a sampled datset with Θ as input. As mentioned earlier, predictions of classifiers on the remaining the training data are used as the high-level features for credit prediction.

Features derived from low-level features: Similar to ngram features, we also propose to extract high-level features by applying a learning algorithm to different sets of the low-level features defined earlier, including demographic features, tweet features, and network features. Differing from high-dimensional ngram features, a bunch of simple

[12]Also known as bootstrapped samples in stacking.

supervised learning algorithms such as Logistic Regression, SVM, and Decision Trees can be used to generate high-level features from these low-level features. In this way, we can derive a number of high-level features from low-level features. These high-level features can further distill hidden evidence in the social data.

For features derived from unigram features, we find that the χ^2 statistics can be as large as 25.76, which pass the significance test. Similar results can be expected for features derived from topic distributions. For features derived from low-level features, these features are already shown to be effective for credit prediction, so we omit a detailed analyses of them. When evaluating the effectiveness of our approach experimentally, we will demonstrate the performance improvement of these high-level features both qualitatively and quantitatively.

6. EXPERIMENTS

In this section, after formally introducing the Weibo dataset for experiments, we evaluate the effectiveness, efficiency, and robustness of our ensemble learning framework for social-data-based personal credit scoring. First, we verify if the ensemble of three categories of low-level features as well as high-level features would improve the overall performance. Second, we compare our Tier-2 classifier GBDT with other state-of-the-art credit scoring algorithms for credit prediction. Third, we evaluate the sensitivity and robustness of our approach. Fourth, we perform empirical case studies to further validate the interpretability of our proposed approach. Last, some limitations of this current study are discussed for clarity and technical soundness. All experiments are performed on a 2.00GHz$\times$ 12 Core CPU, 128GB RAM Standard Server (Windows).

6.1. Weibo DataSet

In this subsection, we first briefly introduce the Weibo platform and data collection methods used to acquire the Weibo dataset, which will serve as the testbed for social-data-based credit scoring. After that, some statistical descriptions of the dataset are presented in detail.

6.1.1. Data Collection. Weibo, the most popular tweet-style social platform in China, has about 600 million registered users as of 2015, among which 76.6 million users are daily active users. It is also reported that 2.8 billion tweets are posted on it each month.[13] In practice, Weibo users' online tweets are publicly available by nature. For example, on the Weibo platform, anyone can access another person's tweet data even if she is not a friend of the corresponding user. The Weibo Open APIs allow us to access and analyze a given user's Weibo data after we are granted privileges by the corresponding user. Generally speaking, Weibo is a very comprehensive, open, and suitable information source for social-data-based credit scoring evaluation.

Specifically, all users in the Weibo dataset have authorized our financial partner to collect their self-disclosed demographics, tweet data, and social networks, which is a common prerequisite to make loans from P2P lending companies. In addition, all identities of these users are anonymized to protect their privacy during our study. In this way, no privacy breaches allow us to study these users' credit risk based on Weibo data. In total, we obtained more than 200,000 users' Weibo data, whose credit labels are already known from our partner's internal data. For the 200,000 users $\mathcal{U} = \{u_i\}_{i=1}^{n}$, most of them are not friends with each other (i.e., not one-hop neighbors in the social network). To make it more difficult, it is impossible for us to download the entire social network due to Weibo's data policy. Fortunately, it is easy to obtain the one-hop relationships of a given user u_i, which can construct the core social network of that

[13] http://expandedramblings.com/index.php/weibo-user-statistics.

Table IX. Statistics of Weibo Dataset Used for Performance Evaluation

Description	Value
# of good credit users	28,830
# of bad credit users	1,507
# of Tweets	6,852,362
# of Tweets by good credit users	6,575,607
# of Tweets by bad credit users	276,755
Total # of words	12,301,485
Size of vocabulary	694,191
Threshold of # of tweets per user	21

user. Therefore, we crawled the ego-network G_i of u_i, which contains all the one-hop connections E_i with respect to u_i and all the neighbor nodes V_i connected to u_i. We denote the total subgraph set by $\mathcal{G} = \{G_i = (V_i, E_i)\}_{i=1}^{n}$. It is worth noting that if we have the privilege to access one's Weibo data, Weibo allows us to crawl that user's one-hop neighbors' basic information (e.g., number of tweets, number of friends, etc.).

6.1.2. Data Description. The Weibo dataset we use contains 7,331,334 tweets, among which 1,912,481 (26.1%) are retweets. The average number of embedded URLs per tweet is 0.184. After removing 4,055 (0.055%) tweets containing only URLs, 7,327,279 tweets are left. 799,835 (11.4%) tweets out of the total contain only mentions or emoticons, with 46,326 (5.8%) containing only mentions and 13,891 (1.7%) containing only emoticons. In total, there are 505,849 different kinds of mentions in the tweets, among which 446,660 (88.3%) have a frequency lower than 5; there are 8,370 different kinds of emoticons in the tweets, among which 6,262 (74.8%) have a frequency lower than 5. The average number of mentions per tweet is 0.234; the average number of emoticons per tweet is 0.4001; and the average number of hashtags per tweet is 0.077. If we remove words with frequency less than 5, 198,935 (28.7%) words are left, while the total number of words in the vocabulary is 694,191.

As described in Section 2.1, a large number of users only post a few tweets on Weibo, but a sufficient number of tweets per user is essential. In the following, we only consider users whose number of tweets is greater than 21 for performance evaluations, which includes 28,830 (95.0%) good credit users and 1,507 (5.0%) bad credit users. For users with tweet numbers between 5 and 20, their datasets are used for training the Tier-1 classifiers. Note that we remove stop words and punctuation from tweets before feature extraction, and we only remove words with frequency less than 5 when applying topic model methods for feature extraction. Detailed statistics after data cleaning and filtering are presented in Table IX.

6.2. Experiment Setup

6.2.1. Evaluation Metrics. To show the effectiveness of the proposed feature sets, we adopt performance evaluation metrics including Precision, Recall, F1-Score, Accuracy, and AUC (Area under the ROC Curve) [Bradley 1997]. To further evaluate our method's performance under the class imbalance setting, we employ the measurement Matthews Correlation Coefficient (MCC) [Matthews 1975], which takes into account both true and false positives and true and false negatives. Similar to other correlation coefficients, MCC ranges from -1 to 1, with 1 representing a perfect prediction and -1 a total disagreement. In addition, we also plot ROC curves and Precision-Recall curves to intuitively compare the predictive performance between implementations and algorithms. When comparing our approach with baselines for credit scoring, we use the same set of evaluation metrics. It is worth mentioning that traditional credit scoring studies usually focus on improving prediction accuracy, which is usually evaluated on balanced

datasets. The adopted evaluation metrics give a more systematic investigation of the credit scoring problem.

6.2.2. Comparison Methodology. To validate the effectiveness of different feature sets, we compare the performance of different experimental settings instantiated by different feature set combinations. For brevity, demographic features, tweet features, network features, and high-level features are abbreviated as DF, TF, NF, and HF, respectively. We also demonstrate the superiority of the Tier-2 learning algorithm GBDT through comparison with other classification algorithms, including Random Forest (RF), Bagging methods (BAG), Naive Bayes (NB), Logistic Regression with L_1-regularization (LR), and SVM with linear kernel (SVM). For RF, BAG, and NB, we use implementations from WEKA [Hall et al. 2009]. For LR, we use the implementation from the package Liblinear,[14] and SVM from LibSVM.[15] Settings and parameters of these baseline algorithms are tuned with grid search in our experiments. For GBDT, we use the implementation from package XGBoost.[16] After parameter selection, we set the learning rate η as 0.1, the maximum tree depth as 3, the number of estimators as 25, and the minimum number of instances for each node as 10 to avoid overfitting. The remaining parameters in the GBDT model are set to default values.

Note that the traditional credit scoring literature has used the above-mentioned machine learning algorithms extensively [Hand and Henley 1997; Crook et al. 2007]. In the selected state-of-the-art baseline algorithms, LR, SVM, and RF have been employed as the main methodology in previous credit scoring studies of Wiginton [1980]; Schebesch and Stecking [2005], and Harris [2013], respectively. Because the dataset is large-scale, only a linear kernel is chosen for baseline SVM. By default, we repeat credit prediction experiments in 10 rounds of 10-fold cross-validation. Usually, we report the average performance $\pm$ a standard deviation of the 10 rounds of executions.

6.3. Performance Comparison with Different Instantiations

In our testbed dataset, the number of bad credit users is significantly smaller than that of good credit ones. To empirically evaluate our approach's performance with respect to decision threshold-sensitive measurements like Precision, Recall, F1-Score, Accuracy, and MCC, we have to deal with the data imbalance issue first. To that end, we first sample 1,500 bad credit users and 1,500 good credit users randomly to construct a balanced dataset. We can then reasonably evaluate our framework's performance with respect to decision threshold-sensitive measurements. For evaluations under the imbalanced setting, we use all the users in our dataset since it represents the real-world case in social-data-based credit scoring. In the following, we will report the performance evaluation results under both the balanced and imbalanced settings. To be more specific, we analyze the performance of low-level and high-level features separately since they are generated in different ways.

6.3.1. Evaluation Under the Balanced Setting. Table X presents the performance of different low-level feature set combinations with respect to Precision, Recall, and F1-Score on the balanced dataset. We can see that different feature combinations have different predictive advantages with respect to different performance measurements. Table XI(a) shows the performance comparison with respect to Accuracy and AUC under the balanced setting. The instantiation with all low-level features as input (i.e., DF+TF+NF) outperforms all other baselines in terms of Accuracy and AUC. Other observations are as follows: (i) tweet features and behavior features are more predictive than network

[14] http://www.csie.ntu.edu.tw/cjlin/liblinear/.
[15] https://www.csie.ntu.edu.tw/cjlin/libsvm/.
[16] https://github.com/tqchen/xgboost/.

Table X. Performance Comparison with Different Instantiations of Our Approach with Respect to Precision, Recall, and F1-Score on the Balanced Dataset

	Good Credit			Bad Credit		
Method	Prec.	Recall	F1-Score	Prec.	Recall	F1-Score
DF	0.5818±0.0043	0.5002±0.0063	0.5370±0.0035	0.5606±0.0025	**0.6346**±0.0065	**0.5955**±0.0052
TF	0.5603±0.0031	0.6047±0.0092	0.5804±0.0051	0.5731±0.0044	0.5298±0.0042	0.5503±0.0080
NF	0.5354±0.0080	0.5204±0.0138	0.5204±0.0056	0.5296±0.0026	0.5484±0.0086	0.5379±0.0072
DF+TF	0.5808±0.0047	**0.6175**±0.0076	**0.5979**±0.0037	0.5899±0.0036	0.5521±0.0086	0.5671±0.0064
DF+NF	0.5647±0.0055	0.5403±0.0084	0.5527±0.0065	0.5592±0.0047	0.5876±0.0076	0.5740±0.0059
TF+NF	0.5690±0.0049	0.5949±0.0098	0.5794±0.0060	0.5768±0.0043	0.5494±0.0079	0.5599±0.0068
DF+TF+NF	**0.5828**±0.0040	0.6113±0.0069	0.5961±0.0055	**0.5920**±0.0050	0.5626±0.0091	0.5757±0.0058

Table XI. Overall Performance Comparison with Different Instantiations of Our Approach with Respect to Accuracy, AUC, and MCC Under Both Balanced and Imbalanced Settings

(a) Balanced Setting

	Overall	
Method	Accuracy	AUC
DF	0.5677±0.0040	0.5880±0.0031
TF	0.5637±0.0062	0.5882±0.0025
NF	0.5317±0.0067	0.5520±0.0050
DF+TF	0.5798±0.0044	0.6125±0.0041
DF+NF	0.5650±0.0040	0.5896±0.0030
TF+NF	0.5758±0.0029	0.6038±0.0020
DF+TF+NF	**0.5856**±0.0029	**0.6203**±0.0042

(b) Imbalanced Setting

	Overall	
Method	AUC	MCC
DF	0.5752±0.0036	0.0452±0.0032
TF	0.5886±0.0042	0.0552±0.0034
NF	0.5733±0.0039	0.0430±0.0028
DF+TF	0.5995±0.0033	0.0676±0.0029
DF+NF	0.5989±0.0020	0.0616±0.0027
TF+NF	0.6107±0.0034	0.0668±0.0046
DF+TF+NF	**0.6166**±0.0041	**0.0736**±0.0024

features; (ii) the progressive integration of DF, TF, and NF shows a tendency of diminishing return to some extent (e.g., the Accuracy improvements from method DF to DF+TF, and DF+TF to DF+TF+NF are 0.121, and 0.058, respectively; (iii) the results also show that combining DF, TF, and NF can achieve the best performance for credit prediction (e.g., DF+TF+NF outperforms DF, TF, and NF methods by 3.16%, 3.89%, and 10.14%, respectively, in terms of Accuracy); and (iv) according to feature importance values returned by instantiation of DF+TF+NF, the most important features include "Age" and "Registration time" from demographic features and "Posting time" and "Sentiment polarity" from tweet features, which is consistent with the feature analysis in Section 5.

6.3.2. Evaluation Under the Imbalanced Setting. To further evaluate the predictive performance of our approach, we present a comparison of different instantiations of our framework on imbalanced datasets. To this end, we construct an imbalanced dataset using all 1,507 bad credit users and 28,830 good credits users. Similar to the balanced setting, 10-fold cross-validation is performed to evaluate the overall performance of different instantiations. When sampling the test dataset, the original imbalance ratio is maintained. For training data, we oversample the positive (minority) class samples (i.e., bad credit users), which is a simplified procedure of SMOTE [Chawla et al. 2002]. Because the imbalance ratio is rather significant in our case, evaluation metrics like Precision, Recall, and Accuracy are not suitable for comparison. Instead, AUC and MCC values of different instantiations of our framework are reported in Table XI(b). It can be seen that (i) among the single feature sets, the best prediction result is achieved by tweet features, which is slightly different from the results under the balanced setting; (ii) the best performance is also achieved when three sets of features are all integrated in terms of both AUC and MCC measurements; and (iii) performance measurement MCC presents similar experimental results to that of AUC, further verifying that DF+TF+NF's performance is the best.

Table XII. Performance Comparison Among Instantiations when Considering High-level Features with Respect to Precision, Recall, and F1-Score on the Balanced Dataset

	Good Credit			Bad Credit		
Method	Prec.	Recall	F1-Score	Prec.	Recall	F1-Score
LF	0.5828±0.0040	**0.6113**±0.0069	0.5961±0.0055	0.5920±0.0050	0.5626±0.0091	0.5757±0.0058
HF	0.5568±0.0046	0.5622±0.0086	0.5559±0.0059	0.5600±0.0042	0.5562±0.0062	0.5538±0.0060
LF+HF	**0.5901**±0.0059	0.6102±0.0069	**0.5966**±0.0051	**0.5991**±0.0058	**0.5672**±0.0058	**0.5851**±0.0047

Table XIII. Overall Performance Comparison Among Instantiations when Considering High-level Features with Respect to Accuracy, AUC, and MCC

(a) Balanced Setting

	Overall	
Method	Accuracy	AUC
LF	0.5856±0.0029	0.6203±0.0042
HF	0.5582±0.0057	0.5765±0.0035
LF+HF	**0.5876**±0.0037	**0.6251**±0.0072

(b) Imbalanced Setting

	Overall	
Method	AUC	MCC
LF	0.6166±0.0041	0.0736±0.0024
HF	0.6268±0.0028	0.0845±0.0018
LF+HF	**0.6375**±0.0016	**0.0900**±0.0017

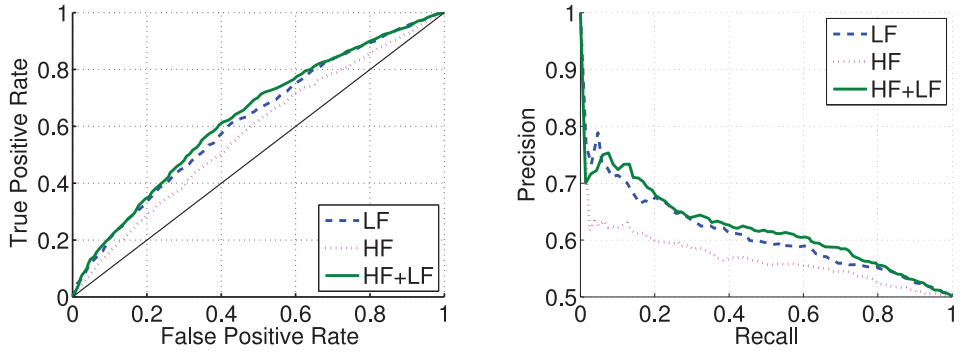
6.3.3. Effects of High-level Features. In the previous evaluations, we only take the low-level features into account. In this subsection, we present the effectiveness evaluation of High-level Features (HF) for credit scoring. The high-level features are generated using the low-level features of those users whose number of tweets is between 5 and 20 (i.e., the Tier-1 classifiers' input data). For convenience of notation, we denote the instantiation DF+TF+NF by LF (i.e., Low-level Features).

Table XII shows the performance comparison with respect to Precision, Recall, and F1-Score among instantiations after taking high-level features into consideration. As we can see, HF+LF outperforms LF in terms of all evaluation measurements except Recall for good credit class. In Table XIII, we further show the overall performance comparison among LF, HF, and LF+HF with respect to Accuracy, AUC, and MCC. Since MCC is more suitable for imbalanced dataset evaluations and Accuracy for balanced dataset evaluations, we use different measurements for different dataset settings in Table XIII. In Table XIII(a), we can see that LF+HF performs better than LF with respect to both Accuracy and AUC under the balanced setting. In Table XIII(b), similar results can be observed in terms of AUC and MCC under the imbalanced setting. Statistical t-test shows that all the performance improvements pass the significance test at a confidence level of 95%. Moreover, the ROC and Precision-Recall curves in Figure 9 show that LF+HF clearly outperforms the baselines, which is coherent with the results in Table XIII. In a word, all these results confirm that it is very effective to take HF into consideration.

Although our work aims to distill informative and discriminative evidence from social data for credit scoring, there can be many factors leading to loan default that are not covered by social data. It is reasonable that the overall performance of social-data-based credit scoring is not very high with these weakly credit-correlated and even noisy social features as input. Despite the fact that our approach's overall AUC value in Table XIII(a) is only 0.625 under the balanced setting, we will show in the case study section that social-data-based credit scoring can be very powerful in terms of interpretability and feasibility.

6.4. Performance Comparison of Learning Algorithms

6.4.1. Effectiveness Evaluation. Table XIV shows the performance comparison among GBDT, RF, BAG, NB, LR, and SVM when they serve as the Tier-2 classifier in our framework. All experiments are evaluated on the balanced dataset with LF+HF as input. The results show that GBDT outperforms all baselines significantly in terms of

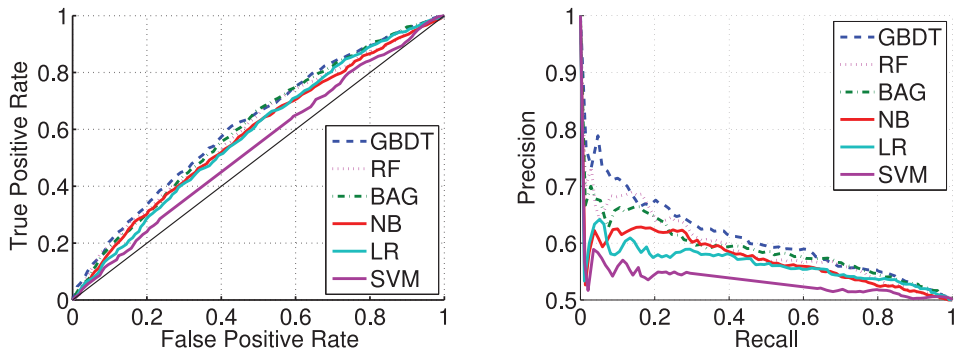


(a) ROC Curve (b) Precision-Recall Curve

Fig. 9. ROC and Precision-Recall curve comparison of different instantiations of our credit model when considering high-level features.

Table XIV. Performance Comparison Between Different Supervised Learning Algorithms on the Balanced Dataset

Algorithm	F1-Score (Bad)	F1-Score (Good)	Accuracy	AUC	Time
GBDT	**0.5966**±0.0051	0.5851±0.0047	**0.5876**±0.0037	**0.6251**±0.0072	**270**s
RF	0.5932±0.0025	0.5333±0.0018	0.5653±0.0018	0.6028±0.0024	345s
BAG	0.5824±0.0019	0.5659±0.0025	0.5743±0.0026	0.6057±0.0023	14341s
NB	0.3502±0.0057	**0.6441**±0.028	0.5402±0.0021	0.5857±0.0027	313s
LR	0.5636±0.0045	0.5665±0.0020	0.5650±0.0033	0.5651±0.0033	1432s
SVM	0.5028±0.0226	0.5593±0.0121	0.5338±0.0025	0.5341±0.0027	41755s

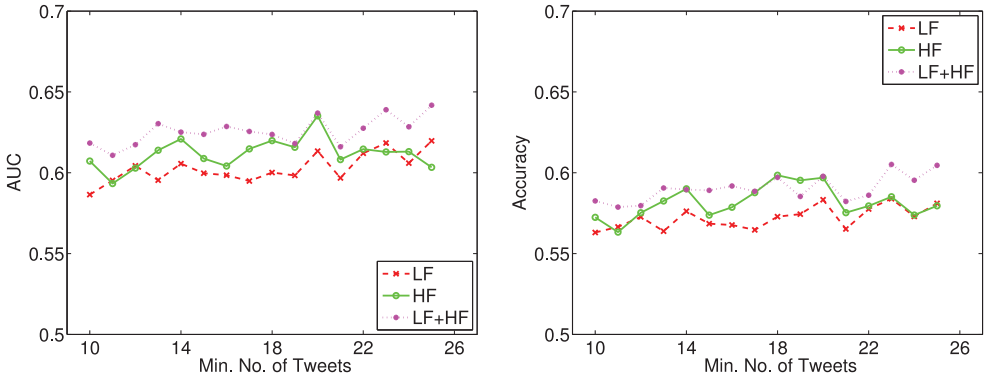


(a) ROC Curve (b) Precision-Recall Curve

Fig. 10. ROC and Precision-Recall curves of different learning algorithms with LF+HF as input.

F1-Score (Bad), AUC, and Accuracy, demonstrating its superior performance for credit scoring as the final classifier. In addition, a statistical t-test shows that the performance differences between GBDT and baselines are significant at the confidence level of 95%. The only exception is NB with respect to F1-Score (Good), but its performance for bad credit users is extremely low. Although the famous SVM method is enhanced with a linear kernel, we can see that GBDT outperforms SVM by 0.0538 (10%) in terms of Accuracy and by 0.091 (17%) in terms of AUC. Figure 10 shows that similar results can be obtained in terms of the ROC and Precision-Recall curves, respectively.

Among the baselines, ensemble methods RF and BAG display considerably better performance than single-model methods such as NB and LR. We can draw the



(a) Accuracy (b) AUC

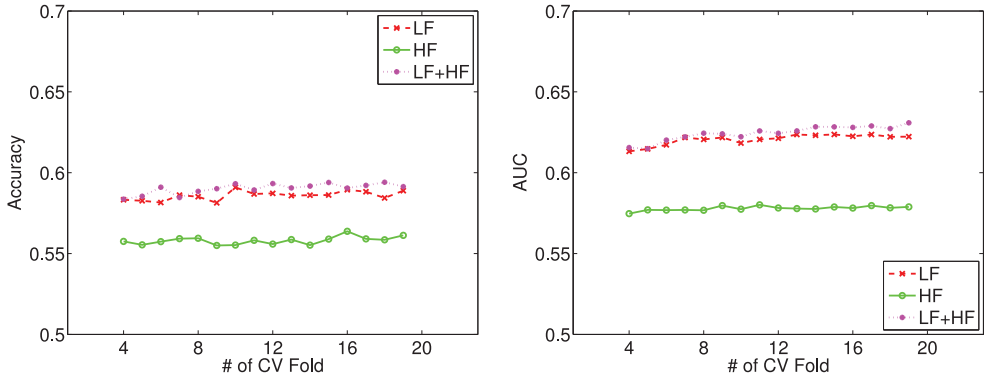
Fig. 11. Robustness of our credit model against varied minimum # of tweets per user.

conclusion that ensemble methods are particularly suitable for learning credit models from a large number of weakly credit-correlated features, which is exactly the situation of our problem. We can also observe that both SVM and LR show extremely low performance. This could be because LR and SVM both require input features to be strong variables and are sensitive to noisy features and/or missing values, rendering them unfit for social-data-based credit scoring. In sum, our ensemble learning framework combining both stacking and boosting strategies is reasonably effective in capturing the small signals hidden in social data.

6.4.2. Efficiency Evaluation. In last column of Table XIV, we report the running time of different learning algorithms. Despite GBDT's outstanding predictive performance, we can see that GBDT is also much more efficient than baseline algorithms. In addition, we note that GBDT is very suitable for parallel implementation, making it very suitable for extremely large-scale social-data-based credit scoring. The BAG algorithm's running time is considerably large because its best performance is achieved when the iteration step is 50 and the percentage of sampled data is 0.8. Even with 20 iterations, the BAG algorithm still takes longer (5112s) than GBDT and the corresponding AUC value decreases to 0.5920. As expected, the SVM algorithm is especially time-consuming since it includes a linear kernel for higher performance. For large-scale datasets, it will cost too much time to use complex kernels (e.g., Gaussian).

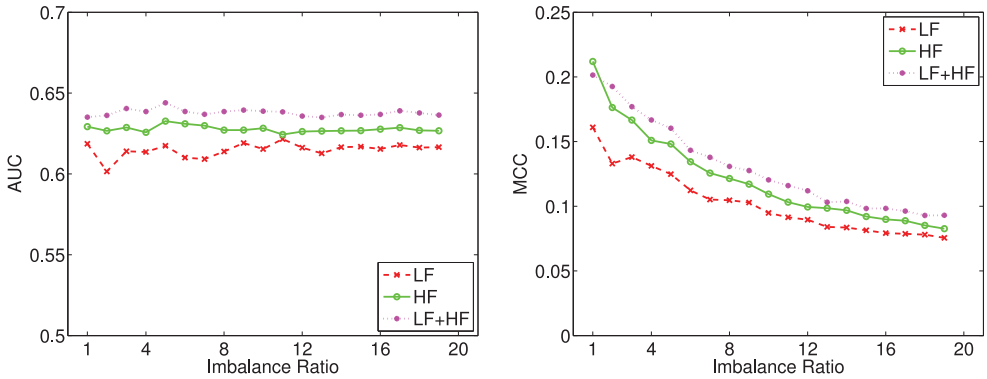
6.5. Model Robustness

6.5.1. Effects of Varied Minimum # of Tweets per User. The threshold of minimum number of tweets per user is set to 21 by default. Here, we examine our method's sensitivity to this threshold parameter. As shown in Figure 3(a), the median number of tweets per user is as few as 2. In this experiment, we use the data of users whose # of tweets is larger than 2 and less than 10 to build Tier-1 classifiers, which further generate high-level features for those users whose # of tweets is larger than 10. In Figure 11, we present the effects of varied minimum # of tweets per user (10∼25) on the performance of credit prediction. From these curves, we can see that it is appropriate to set the threshold to 21 in our experiments. It is also clear that as the minimum # of tweets per user increases, the predictive performance increases to some extent with respect to Accuracy and AUC. From the slightly growing trend of the curves, we can expect that the performance of our credit model will further improve as the threshold continues to increase.



(a) Accuracy (b) AUC

Fig. 12. Robustness of our credit model against varied # of folds for cross-validation on the balanced dataset.



(a) AUC (b) MCC

Fig. 13. Robustness of our credit model against varied imbalance ratios between good and bad credit users.

6.5.2. Effects of Varied Cross-validation Folds. We examine the robustness of our credit model in terms of number of folds for cross-validation in Figure 12. Experiments are performed on the balanced dataset mentioned earlier. It is worth noting that varying the number of cross-validation folds equals changing the training data ratios during performance evaluations. Figure 12 shows that as the number of cross-validation folds increases from 4 to 19, the overall performance of our credit model remains almost unchanged in terms of both Accuracy and AUC. This phenomenon also justifies our setting of the cross-validation fold number. From another perspective, we can also say that the prepared Weibo dataset is already large enough for evaluating the performance of our social-data-based credit model.

6.5.3. Effects of Varied Data Imbalance Ratios. The data imbalance issue is particularly important for social-data-based personal credit scoring. We present the effects of varied imbalance ratios over our credit model's performance in Figure 13. The experiments are performed with users whose number of tweets is larger than 21. We can see that when the imbalance ratio increases in Figure 13(a), the predictive performance remains almost unchanged in terms of AUC. This phenomenon clearly demonstrates that our model is capable of dealing with the data imbalance issue. The decreasing trend of curves in Figure 13(b) may be because the number of bad credit users is too small in

our dataset, and MCC is not a good measurement for extremely imbalanced dataset. On the contrary, AUC is insensitive to decision thresholds and therefore remains stable when the imbalance ratio increases.

In summary, the preceding parameter sensitivity studies demonstrate that our model is rather stable under different settings. The robustness of our credit scoring model is satisfactory for different real-world cases.

6.6. Case Studies

In the subsection, we perform case studies to show our approach's interpretability and feasibility, which are also essential requirements for real-world credit scoring systems in deployment.

6.6.1. Prediction on Single Users. We first train our credit model with the balanced dataset as mentioned earlier, and we predict the class probability of a randomly selected test dataset composed of 14 bad credit users and 14 good credit users. Then, we rank the 28 users according to estimated class probabilities. Luckily, the two top-ranked users of both good and bad credit classes are all correctly labeled. In the following, the two top-ranked good credit users and two top-ranked bad credit users are chosen for detailed analyses. We stop trying larger numbers of case study users when we find both top-ranked good and bad credit users are correctly labeled.

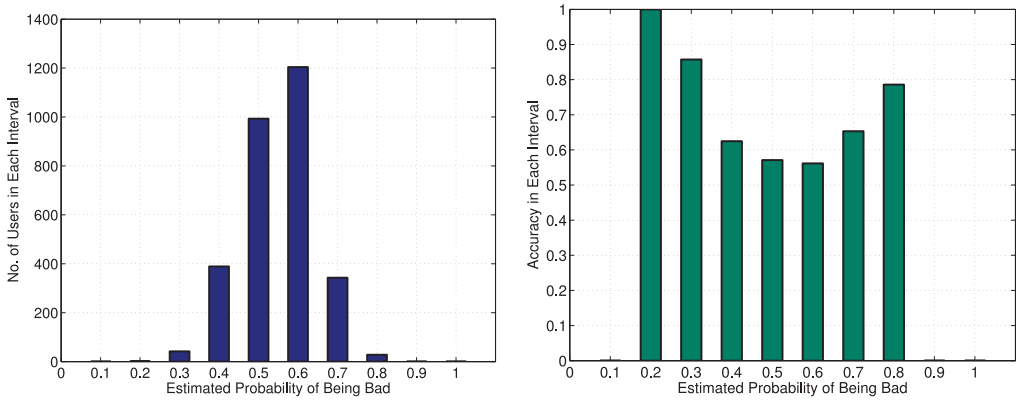
With regard to the two good credit users, we find that (i) the two good credit users tend to post relatively short tweets, mostly about their everyday lives; (ii) one of them shows signs indicating that he is a skilled computer programmer; (iii) the other appears to be a college teacher and enjoys posting very short tweets about personal thoughts and feelings; and (iv) the topics of these short tweets range from sports, political news, films to pop stars.

With regard to the two bad credit users, we find that (i) one of them frequently shares famous quotes on love, friendships and so on, and posts a lot about horoscopes; and (ii) the other bad credit user seems to be an online retailer who treats Weibo as an advertising platform, frequently tweeting about online shopping, commercial campaigns, and featured products from her online store. We also observe that both bad credit users are youngsters who have just begun their careers.

In a word, these case studies show that good credit users are more likely to tweet about their everyday lives and share with their friends on social platforms, while bad credit users tend to take advantage of these platforms to seek entertainment or commercial opportunities. We can reckon that good credit users tweet more creatively and are more likely to be professional workers, which is in accord with the intuitive insights we presented earlier.

6.6.2. Prediction at Different Probability Intervals. In Figure 14, we show the performance of our approach by comparing its predictive ability at different probability intervals. As we all know, the estimated probability of being in a certain class for each test user can be obtained from classification models. We use all 3,000 users' estimated probability of being bad credit risks, which can be obtained with cross-validation procedures, to plot these figures. Specifically, we divide the range of estimated probabilities (0∼1) into 10 equal intervals, and each interval has a length of 0.1.

In Figure 14(a), we can see that the distribution of users in each estimated probability interval follows an overall Gaussian curve. Although most users' estimated probabilities lie around 0.5, there still exist a number of users whose estimated probabilities are close to 0 or 1. In Figure 14(b), we show the prediction accuracy computed for users in each interval. It can be observed that although prediction accuracy for intervals around 0.5 is considerably low, the accuracy values for intervals far from 0.5 are very high. These results are coherent with our intuitions of estimated probabilities



(a) No. of Users (b) Predicted Accuracy

Fig. 14. The distribution of users with respect to estimated probability of being bad credit risks and the corresponding accuracy values.

output by classification models. On the other hand, they also coincide with the phenomenon that the top-ranked two users in the 14 selected good/bad credit users are all correctly labeled by our credit model.

Empirically, we find that, in the case of Figure 14, users located in two tail parts have a prediction accuracy above 0.78. Thus, if only users with estimated probabilities far from 0.5 are considered, we can have very good predictive performance for credit scoring. For users who are hard to classify, additional manual inspection can then play an important role in credit scoring. In this way, our social-data-based credit scoring approach can filter out easy cases and save costs in understanding customers' creditworthiness. In short, we can conclude that our credit scoring approach can be very useful when used properly even though its overall performance is limited.

6.7. Limitations and Discussions

Although this study attempts to distill informative and discriminative evidence from social data for personal credit scoring, we acknowledge that it is very hard to predict one's default risk singly based on social data.

Here, we'd like to discuss some possible limitations of our social-data-based study to evaluate personal credit risk. First, this study only predicts whether a user's credit class is "good" or "bad," which is just the first step in generating the final credit score. Although this is a convention in credit scoring literature, post-processing steps are needed to accurately decide a person's credit score. Second, any given dataset carries biases to at least some extent, and our Weibo dataset is no exception. For example, as shown in Figure 2(a), young adults aged between 20 and 40 are disproportionately presented in the testbed. It is also worth noting that some evaluations of features in Section 5 could be false positive and only work when combined due to small significance values. Third, our investigation is mainly conducted under the micro-blogging setting of Weibo. Some of the results might not generalize to other types of social platforms like Facebook, LinkedIn, and the like. In addition, personal characteristics on the social web can differacross cultures, and credit scoring regulations may vary greatly across countries [Thomas et al. 2002]. Fourth, although one's online high-quality social connections are hard to manipulate, the credibility of online tweets or personal profiles can be undermined by malicious users [Gao et al. 2010]. Similar to traditional user survey data, a thorough social data trustworthiness check should also be done before

credit evaluation. Last, the setting of appropriate credit class definitio (i.e., default definition) varies in practice. For example, one's creditworthiness can change from "good" to "bad" very frequently. It is not easy to conclude that one is "good credit" or not. In this work, we only follow the most common and popular definition recommended by our financial partner.

However, as mentioned in Section 1, our work for social-data-based credit scoring can at least be viewed as a complement to existing credit scoring work. For example, our work is particularly valuable for "unbanked" or "unscorable" consumers. It is also worth noting that the casual relationships between the observations and the credit labels are not very clear in the research community to date. There could be other factors that lead to loan default not covered by social data. It is therefore reasonable that the overall performance of social-data-based credit scoring is not very high. Since credit is an especially complex and subtle attribute of individuals in real life, even existing credit scoring methods cannot guarantee very high performance. For example, based on a real-world loan dataset, the obtained AUC value of credit default prediction in Harris [2013] is only around 65%, even though a random forest algorithm [Breiman 2001] is employed as the classifier and the input includes 18 well-known traditional credit-related user attributes. In real-world practice, even traditional credit scoring methods' prediction results are not always directly usable. To obtain practical credit risk evaluations and high-quality credit scoring, additional human judgments of inferred credit labels are usually needed.

On the other hand, similar to other web data mining [van Wel and Royakkers 2004] or big data problems [Boyd and Crawford 2012], ethical issues exist in social-data-based credit scoring. Important parts of our future work, as described at the end of this article, are to study how to adapt our proposed method to other social platforms with appropriate privacy-preserving strategies and how to deal with the effects of different credit class definitions. With increasingly more available online social data, the development of data sharing policies, and the improvement of social data quality, we believe that social data will become as important as financial institutions' internal data in the near future. For example, in the previous subsection, our approach shows fairly good performance for users located in the tail parts of Figure 14.

7. RELATED WORK

Because we aim to profile users' credit attributes from social data, we examine related work in both traditional consumer credit scoring and user profiling on social media. Traditional consumer credit scoring literature mainly focuses on consumer credit scoring, which is usually focused on small loans applied for by individuals, the same subject as our work. On the other hand, social-data-based credit scoring can be viewed as estimating the specific user attribute of creditworthiness of individuals from social data, which is closely related to the task of user profiling on social media.

7.1. Traditional Consumer Credit Scoring

A wealth of research has been conducted on credit risk management and consumer credit scoring [Baesens et al. 2003; Rosenberg and Gleit 1994; Hand and Henley 1997; Thomas et al. 2002; Crook et al. 2007] during the past few decades. Previous studies apply a variety of statistical methods, including discriminant analysis [Eisenbeis 1978], support vector machine [Schebesch and Stecking 2005], logistic regression [Wiginton 1980], decision tree [Arminger et al. 1997], neural networks [Jensen 1992], k-nearest neighbors [Henley and Hand 1996], time varying models [Frydman et al. 1985], and genetic algorithms [Ong et al. 2005] for better personal credit scoring. Recent years have also witnessed the fast development of more advanced statistical learning methods applied for credit scoring [Huang et al. 2007; Hsieh and Hung 2010; Yap et al. 2011;

Kruppa et al. 2013; Harris 2015; Kozeny 2015], which mainly rely on the state-of-the-art algorithms developed in data mining and machine learning. In particular, Harris [2013] assesses individuals' credit default risk by exploring the optimal default definition selection algorithm, which tries to select the best default definition for building models.

In addition to statistical methods for building credit scoring models, there is also a body of work aiming at the analysis of specific factors in credit risk assessment. Vissing-Jorgensen [2011] presents the determinant of consumer credit default from retail chain store datasets and finds that the products a consumer purchases provide substantial information about potential default risk. Chatterjee et al. [2007] study the general equilibrium of an economy with unsecured loans and characterize the circumstances under which defaults happen. Einav and Jonathan [Adams et al. 2007] observe that default rates rise significantly with loan size based on data from a large auto sales company serving the subprime market. Agarwal et al. [2008]'s study based on a proprietary panel dataset from a large US bank shows that credit borrowers who are more experienced, high-income, and middle-aged learn better from negative feedback. Agarwal et al. [2009] also examine users' payday loan creditworthiness with credit scores from FICO. Karlan and Zinman [2009]'s study shows that 7% to 16% of defaults are due to asymmetric information problems through a new field experiment.

There are two key differences between our work and the traditional consumer credit scoring studies. First, we focus on social data for extracting evidence about consumers' credit risk, while the traditional models are based on transactional loan/payment records, credit reports, or demographic survey data and so on, most of which are strongly credit-correlated data sources. Second, we propose to capture the weak signals pertinent to credit risk in social data by an ensemble learning framework, instead of using simple statistical methods as in existing works.

7.2. User Profiling on Social Media

User profiling from social data has not been studied extensively until the recent boom of social media platforms like Twitter, Facebook, Weibo and the like. Rao et al. [2010] first attempted to classify user attributes including gender, age, region, and political affiliation based on ngram features and sociolinguistic features from users' tweets. Pennacchiotti and Popescu [2011] conducted a user attribute profiling study for attributes including political affiliation and affinity to a certain brand with more diverse social features. Cheng et al. [2010] identified "local" words from training tweets first and then inferred users' home locations based on these local words.

Other studies inferring user attributes like gender [Burger et al. 2011; Fink et al. 2012; Bergsma and Durme 2013], age [Rosenthal and McKeown 2011; Nguyen et al. 2013; Goswami et al. 2009; Peersman et al. 2011], and the like also take advantage of tweet content. Zhong et al. [2015] infer user demographic attributes such as gender, age, education background, and marital status singly based on user-generated location check-ins on social media. Chen et al. [2014a] propose to apply frequent sequential pattern mining techniques to model users' mobility profiles from their historical visit sequences. Aside from user-generated content, social connections between online users are also explored for user attribute inference in Mislove et al. [2010], Backstrom et al. [2010], Zeng et al. [2013], Dong et al. [2014], and Chen et al. [2014b]. Taking one step further, Li et al. [2012] propose a unified discriminative and probabilistic framework that captures both social network and user-centric data for inferring users' home location. They also propose a user co-profiling methodology [Li and Chang 2014] that simultaneously models user relationship types and attributes for user profiling.

The most critical difference between our work and previous user profiling studies is that our study is an integration of various weakly credit-correlated features, which include basic user demographic information such as age, gender, and education, as

well as a large number of tweet and network features. Second, our work is purposed for inferring the specific attribute of user creditworthiness and is performed under guiding principles from traditional credit scoring literature and insights from empirical observations. In our previous study [Guo et al. 2016], we also directly employed social data for personal credit scoring, but only topic model techniques were used to extract latent user behavior dimensions for credit scoring. Because topic models usually require lots of time to infer and predict, the method in Guo et al. [2016] is only suitable for small datasets. For example, demographic, tweet, and network data are all utilized in this work, whereas only tweet data are taken into consideration in Guo et al. [2016]. What's more, experiments in the two studies are performed on different datasets, which are collected at different time periods and of different sizes.

8. CONCLUSION

In the big social data era, astonishing improvements in credit quality and customer performance have been achieved by combining social media data with real-life financial data. In this article, the problem of heterogeneous social data-based personal credit scoring is formally introduced. To systematically study the problem, we use Weibo, one of the most typical micro-blogging and social-networking platforms, as the testbed for a data-driven investigation of relationships between user credit labels and social data.

Inspired by research in traditional consumer credit scoring, we propose practical principles and insights that guide our feature extraction process. Specifically, we extract three different categories of low-level features from Weibo. Detailed analyses and discussions are performed for these low-level prediction features. In addition, high-level features exploiting the stacking technique are also proposed to further enhance our framework's prediction performance. Taking both low- and high-level features into account as input, we are able to harness the most information available in the social data for credit scoring. To handle these diverse sets of prediction features, we propose a unified two-tier credit scoring model armed with both boosting and stacking strategies. Extensive experiments on real-world social data show that our framework is effective, efficient, and robust in harnessing the weakly credit-correlated signals embedded in social data. Our work is especially practical for those with scant credit records who might otherwise have trouble receiving loans. While our study focuses on data analysis and the overall performance is not very high, we believe that it opens doors to building high-quality social-data-based credit scoring systems.

In the future, we plan to explore our social-data-based credit scoring framework using different types of social datasets, such as datasets from location-based social networks. We also want to study the effects of different credit class definitions on personal credit scoring when using social data as input. Another future direction is to design more effective features and methods for mining heterogeneous social data. We are also very interested in studying our framework's performance when both social data and traditional financial data are available. These future works can help us have a better understanding of the possibilities and limits of social-data-based credit scoring.

ACKNOWLEDGMENTS

The authors thank Juan Du and Runquan Xie for their help in data preparation and anonymous reviewers for their constructive comments.

REFERENCES

William Adams, Liran Einav, and Jonathan Levin. 2007. *Liquidity Constraints and Imperfect Information in Subprime Lending*. Technical Report. National Bureau of Economic Research.

Sumit Agarwal, John C. Driscoll, Xavier Gabaix, and David Laibson. 2008. *Learning in the Credit Card Market*. Technical Report. National Bureau of Economic Research.

Sumit Agarwal, Paige M. Skiba, and Jeremy Tobacman. 2009. *Payday Loans and Credit Cards: New Liquidity and Credit Scoring Puzzles?* Technical Report. National Bureau of Economic Research.

Gerhard Arminger, Daniel Enache, and Thorsten Bonne. 1997. Analyzing credit risk data: A comparison of logistic discrimination, classification tree analysis, and feedforward networks. *Computational Statistics* 12, 2 (1997).

Alexander Bachmann, Alexander Becker, Daniel Buerckner, Michel Hilker, Frank Kock, Mark Lehmann, Phillip Tiburtius, and Burkhardt Funk. 2011. Online peer-to-peer lending – a literature review. *Journal of Internet Banking and Commerce* 16, 2 (2011), 1.

Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find me if you can: Improving geographical prediction with social and spatial proximity. In *WWW*. 61–70.

Bart Baesens, Tony Van Gestel, Stijn Viaene, Maria Stepanova, Johan Suykens, and Jan Vanthienen. 2003. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society* 54, 6 (2003), 627–635.

Shane Bergsma and Benjamin Van Durme. 2013. Using conceptual class attributes to characterize social media users. *ACL (1)*. 710–720.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

Andreas Blochlinger and Markus Leippold. 2006. Economic benefit of powerful credit scoring. *Journal of Banking and Finance* 30, 3 (2006), 851–873.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2, 1 (2011), 1–8.

Danah Boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15, 5 (2012), 662–679.

Andrew P. Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30, 7 (1997), 1145–1159.

Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32.

John D. Burger, John C. Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *EMNLP*. 1301–1309.

Moses S. Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *STOC*. ACM, 380–388.

Satyajit Chatterjee, Dean Corbae, Makoto Nakajima, and José-Víctor Ríos-Rull. 2007. A quantitative theory of unsecured consumer credit with risk of default. *Econometrica* 75, 6 (2007), 1525–1589.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* (2002), 321–357.

Xihui Chen, Jun Pang, and Ran Xue. 2014a. Constructing and comparing user mobility profiles. *ACM Transactions on the Web* 8, 4, Article 21 (Nov. 2014), 25 pages.

Zhuohua Chen, Feida Zhu, Guangming Guo, and Hongyan Liu. 2014b. User profiling via affinity-aware friendship network. In *Social Informatics*. Springer, 151–165.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: A content-based approach to geo-locating twitter users. In *CIKM*. 759–768.

Jonathan N. Crook, David B. Edelman, and Lyn C. Thomas. 2007. Recent developments in consumer credit risk assessment. *European Journal of Operational Research* 183, 3 (2007), 1447–1465.

Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, and Nitesh V. Chawla. 2014. Inferring user demographics and social strategies in mobile social networks. In *KDD*. 15–24.

Liran Einav, Mark Jenkins, and Jonathan Levin. 2013. The impact of credit scoring on consumer lending. *The RAND Journal of Economics* 44, 2 (2013), 249–274.

Robert A. Eisenbeis. 1978. Problems in applying discriminant analysis in credit scoring models. *Journal of Banking & Finance* 2, 3 (1978), 205–219.

Clayton Fink, Jonathon Kopecky, and Maksym Morawski. 2012. Inferring gender from the content of tweets: A region specific example. In *ICWSM*.

Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* (2001), 1189–1232.

Halina Frydman, Jarl G. Kallberg, and Duen-Li Kao. 1985. Testing the adequacy of Markov chain and mover-stayer models as representations of credit behavior. *Operations Research* 33, 6 (1985), 1203–1214.

Hongyu Gao, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen, and Ben Y. Zhao. 2010. Detecting and characterizing social spam campaigns. In *IMC*. ACM, 35–47.

Elizabeth M. Gerber and Julie Hui. 2013. Crowdfunding: Motivations and deterrents for participation. *ACM Transactions on Computer-Human Interaction* 20, 6 (2013), 34.

Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric analysis of bloggers' age and gender. In *ICWSM*.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101, suppl 1 (2004), 5228–5235.

Guangming Guo, Feida Zhu, Enhong Chen, Le Wu, Qi Liu, Yingling Liu, and Minghui Qiu. 2016. Personal credit profiling via latent user behavior dimensions on social media. In *PAKDD 2016*. 130–142.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations Newsletter* 11, 1 (Nov. 2009), 10–18.

David J. Hand and William E. Henley. 1997. Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 160, 3 (1997), 523–541.

Terry Harris. 2013. Default definition selection for credit scoring. *Artificial Intelligence Research* 2, 4 (2013), p49.

Terry Harris. 2015. Credit scoring using the clustered support vector machine. *Expert Systems with Applications* 42, 2 (2015), 741–750.

W. E. Henley and David J. Hand. 1996. A k-nearest-neighbour classifier for assessing consumer credit risk. *The Statistician* (1996), 77–95.

Liangjie Hong and Brian D. Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the 1st Workshop on Social Media Analytics*. ACM, 80–88.

Nan-Chen Hsieh and Lun-Ping Hung. 2010. A data driven ensemble classifier for credit scoring analysis. *Expert Systems with Applications* 37, 1 (2010), 534–545.

Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou. 2015. Short text understanding through lexical-semantic analysis. In *ICDE*. 495–506.

Cheng-Lung Huang, Mu-Chen Chen, and Chieh-Jen Wang. 2007. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications* 33, 4 (2007), 847–856.

Michael K. Hulme and Collette Wright. 2006. Internet based social lending: Past, present and future. *Social Futures Observatory* 11 (2006), 1–115.

Akshay Java, Xiaodan Song, Tim Finin, and Belle L. Tseng. 2007. Why we twitter: An analysis of a microblogging community. In *WebKDD/SNA-KDD*. 118–138.

Herbert L. Jensen. 1992. Using neural networks for credit scoring. *Managerial Finance* 18, 6 (1992), 15–26.

Dean Karlan and Jonathan Zinman. 2009. Observing unobservables: Identifying information asymmetries with a consumer credit field experiment. *Econometrica* 77, 6 (2009), 1993–2008.

David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *KDD*. ACM, 137–146.

Vaclav Kozeny. 2015. Genetic algorithms for credit scoring. *Expert Systems with Applications* 42, 6 (April 2015), 2998–3004.

Jochen Kruppa, Alexandra Schwarz, Gerhard Arminger, and Andreas Ziegler. 2013. Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications* 40, 13 (2013), 5125–5131.

Rui Li and Chi Wang Kevin Chen-Chuan Chang. 2014. User profiling in an ego network: Co-profiling attributes and relationships. In *WWW*.

Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. 2012. Towards social user profiling: Unified and discriminative influence model for inferring home locations. In *KDD*. 1023–1031.

Brian W. Matthews. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405, 2 (1975), 442–451.

Alan Mislove, Bimal Viswanath, P. Krishna Gummadi, and Peter Druschel. 2010. You are who you know: Inferring user profiles in online social networks. In *WSDM*. 251–260.

Ethan Mollick. 2014. The dynamics of crowdfunding: An exploratory study. *Journal of Business Venturing* 29, 1 (2014), 1–16.

Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. "How old do you think I am?" A study of language and age in twitter. In *ICWSM*.

Chorng-Shyong Ong, Jih-Jeng Huang, and Gwo-Hshiung Tzeng. 2005. Building credit scoring models using genetic programming. *Expert Systems with Applications* 29, 1 (2005), 41–47.

Michael J. Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *ICWSM*. 265–272.

Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *SMUC*. 37–44.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. Democrats, Republicans and Starbucks afficionados: User classification in twitter. In *KDD*. 430–438.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *SMUC*. 37–44.

Lior Rokach. 2010. Ensemble-based classifiers. *Artificial Intelligence Review* 33, 1–2 (2010), 1–39.

Eric Rosenberg and Alan Gleit. 1994. Quantitative methods in credit management: A survey. *Operations Research* 42, 4 (1994), 589–613.

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In *ACL*. 763–772.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *WWW*. ACM, 851–860.

Klaus B. Schebesch and Ralf Stecking. 2005. Support vector machines for classifying and describing credit applicants: Detecting typical and critical regions. *Journal of the Operational Research Society* 56, 9 (2005), 1082–1088.

Lyn C. Thomas, David B. Edelman, and Jonathan N. Crook. 2002. *Credit Scoring and Its Applications*. SIAM.

Lita van Wel and Lambèr Royakkers. 2004. Ethical issues in web data mining. *Ethics and Information Technology* 6, 2 (2004), 129–140.

Annette Vissing-Jorgensen. 2011. Consumer credit: Learning your customer's default risk from what (s)he buys. Available at SSRN: http://ssrn.com/abstract=2023238 (2011).

John C. Wiginton. 1980. A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis* 15, 03 (1980), 757–770.

Bing Xiang and Liang Zhou. 2014. Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. In *ACL*. 434–439.

Bee Wah Yap, Seng Huat Ong, and Nor Huselina Mohamed Husain. 2011. Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems and Applications* 38, 10 (2011), 13274–13283.

Guangxiang Zeng, Ping Luo, Enhong Chen, and Min Wang. 2013. From social user activities to people affiliation. In *ICDM*.

Hongke Zhao, Qi Liu, Guifeng Wang, Yong Ge, and Enhong Chen. 2016. Portfolio selections in P2P lending: A multi-objective perspective. In *KDD (KDD'16)*. ACM, 2075–2084.

Yuan Zhong, Nicholas Jing Yuan, Wen Zhong, Fuzheng Zhang, and Xing Xie. 2015. You are where you go: Inferring demographic attributes from location check-ins. In *WSDM*. 295–304.

Received December 2015; revised June 2016; accepted September 2016