Cyber-Social-Physical Features for Mood Prediction over Online Social Networks

Chaima DHAHRI Kazunori MATSUMOTO and Keiichiro HOASHI

KDDI Research, Inc 2-1-15 Ohara, Fujimino-shi, Saitama, 356-8502 Japan

E-mail: {ch-dhahri, matsu, hoashi}@kddi-research.jp

Abstract Context-Aware Recommendation Systems (CARS) are more effective when adapting their recommendations to a specific user preference. Since modal context (mood) has a direct impact on user preferences, we aim at having an accurate mood prediction to improve recommendation performance. Online social networks (OSNs) have grown rapidly over the last decade. These social platforms provide the opportunity to gather the distributed online activities for each user. Tracking and aggregating these data could result in useful insights for user modeling and understanding. In this paper, we built a personalized system that can predict the upcoming user mood even in days without text-type tweets. We, first, studied the correlation of three types of features (cyber, social and physical) with a user mood. Then, used these features to train a predictive system. The results suggest a statistically significant correlation between user mood and his cyber, social and physical activities distributed among different OSNs which leads to a low RMSE in our predictive system.

Keyword Mood prediction, Recommendation system, OSNs analysis, Pearson correlation

1. Introduction

With the appearance of context-aware recommendation systems, researchers gave more attention to user-context analysis in order to provide a more accurate recommendation. In such system, learning user preference is a crucial step. This learning step can be done explicitly by asking the user directly or implicitly by tracking its behavior. With the growth of online social networks (OSNs) and the increasing number of active users, online activity behavior become a rich source of real-time data from which user preference can be learned.

User preference may vary with context such as location, time-of-day/ day-of-week, mood, weather/season, people being with, etc. Tracking this variation in real time is a hard task which needs annotation from users. To overcome the difficulty of this task, various methods in the literature have been proposed to analyze user online data, extract daily patterns and track a variation or a change in these patterns.

Statistics from Global Web Index show that the average number of social media accounts per user is 5.54 where people are actively using 2.82 social accounts. By posting on these accounts, users express their likes, attitude, activities, etc. resulting in a huge volume of data distributed cross OSNs. Despite its abundance, this data has an obvious user-centric characteristic which make it more complicated.

In this work, we focus at predicting the mood/ change of mood for each user by analyzing its cross-OSN activities on different dimensions. What we mean by mood, is the variation in the feeling of a person. This is different from personality traits which is more stable over time. The time variety characteristics makes it more difficult to infer or predict.

Existing works focused on analyzing the content of a user post (usually tweets from Twitter) by extracting linguistic and psychological features and/or features related to cyber activities (count of positive/negative tweets, count of active/passive posts, time of post, etc.). However, users are using different types of social networks that generate data with multiple dimensions on cyber, social and physical spaces. The cyber space refers to the user online activity. The social space is related to the user social connectivity, people the user is going out with, etc. The physical space shows the activity and/or the location (restaurant, for example) in real world.

Having multiple accounts, a user can post on one account or on several accounts, can post either text-only, image/link/hashtag –only or multi-modal (text and image/video) posts. Let's consider the scenario where the user posts a non-text tweet on Twitter or shares its location from Foursquare. Conventional works can not predict the mood because, no text is

available to analyze, in this case. However, in our method, we rely, not only on cyber features but also on features related to physical (check-in) and social (mentioned people) spaces. These features allow us to predict the mood even when the user is not tweeting that day by looking at his location or people he is with.

To summarize, we proposed the use of cyber, social and physical features to train a personalized mood predictive system that detects the upcoming mood of a user in days where a text-type tweets is not available and a sentiment classification is not possible.

2. Related works

Many works in the literature analyze the user online data to predict its psychological state. These works focused on:

- A more stable psychological state like personality traits of a user (OCEAN) [1,2]. To do so, the authors in [1] studied the correlation between OCEAN and the type of Twitter user (Listener, Popular, Highly-read and influential). However, in [2], the correlation was calculated between OCEAN and some features extracted from Instagram images.
- Psychological illness detection like depressive disorder [3] or postpartum depression in new mothers [4]. The authors use features related to emotion, engagement, ego-network and linguistic style to classifier users into depressive and non- depressive classes [3] and to detect the change in emotions in new mothers [4].
- Mood or change of mood detection like the work in [5,6,7]. A collective sentiment study was considered in [5] where the authors compare sentiment change over multiple topics (iPhone, Android, Blackberry). [6] explores a number of potential features related to mood; linguistic, psychological features (LIWC), gender, diurnal online activity and personal activity (LIWC). No results were shown related to correlation or prediction. A more recent work is in [7] where the authors extracted features based on online activity on Twitter and Facebook to track the change of mood of a user. However, they collected a very specific dataset (32 highly active students) in a specific period of time (exam then vacation). The data was aggregated using a window of 7 days and results show that only 16 students was highly correlated.

In the papers mentioned above, except the work in [1,2] which focus on personality trait prediction, the main extracted features assume the existence of **text-based** content in a tweet to extract psychological and linguistic features, which is not always the case. Previous works are not effective in cases where a user posts on Twitter without writing any text-based content:

- A user can mention a friend having lunch with at a specific restaurant, for example.
- A user share its post from his other OSN accounts (Foursquare, Instagram, Fitbit, etc.). For example, he can share its check-in from Foursquare (link-based context).

Besides, in conventional approaches, the mood was aggregated on a time period of 7 days which is considered long considering the characteristics of the mood. Moreover, most of the previous works analyzed a small number of users in a specific period of time which would lose the generality of their proposal.

To overcome this, we opt for:

- Extracting 3 types of features not related to linguistic type of features: cyber, social and physical (location). These features will be explained in the next section.
- Aggregating the mood on a shorter period of time (1 day): we extracted features aggregated on a specific day to predict the mood of the following day.
- Using an available online dataset (unlike [7] who considered a specific type of users on specific period of time).

3. Proposal

3.1. Methodology

In order to predict the upcoming mood of a user, we build a multi-stage system as shown in Figure 1. First, for each user, we downloaded its tweets using twitter search API. These tweets were, then, classified into positive and negative sentiments and

aggregated per day. This classification will be used to determine the mood of user per day by calculating a ratio of mood (a value between 0 and 1):

$$mood = \frac{number\ of\ positive\ tweets}{number\ of\ positive+negative\ tweets} \quad (1)$$

We extracted cyber, social and physical types of features from each tweet. Then, we calculated the correlation (Pearson coefficient) between these features and the *mood* per day. The result of the correlation will be used to train the predictive system.

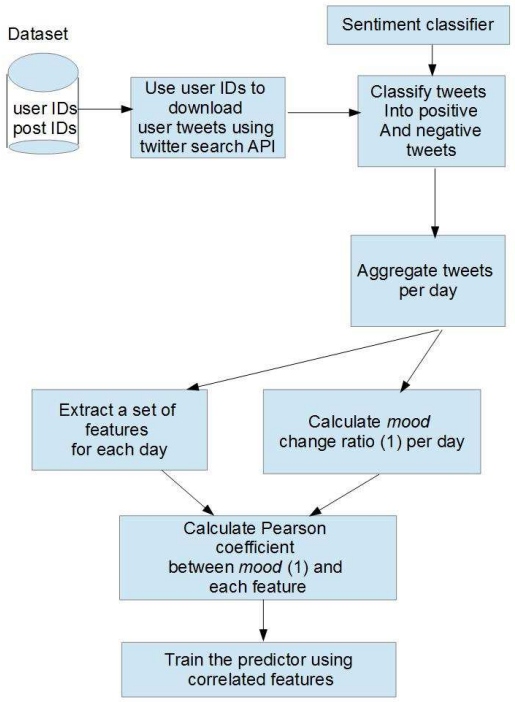


Figure 1: Flow chart of our mood prediction system

3.2. Data Collection

3.2.1. Dataset 1

This dataset is only used to build our sentiment classifier. We used smileys in our request to Twitter API to download positive (☺) and negative (☹) tweets. We downloaded 6451 negative tweets and 7000 positive tweets.

3.2.2. Dataset 2

This dataset is used to predict the user mood. We used an available online dataset having 850 users' IDs, post IDs on three OSNs (twitter, Foursquare and Instagram) [8]. After filtering some corrupted users, we kept the IDs of 800 users that we will use to download recent tweets per user. For now, we are using only Twitter account ID of each user to download data (tweets). However, the tweet's source could be either from Twitter or from other social accounts. We retrieved up to 2000 tweets per user from Jan. 2017. The tweets contain user tweets and retweets. The average number of tweets per user per day is 13.45 tweets. Since the users in this dataset are mostly positive users, we focused, in the training phase, on days where negative and positive tweets exist. In our dataset, we found on average 55 days per user where both negative and positive tweets exist. These days are not necessarily consecutive.

3.3. Feature Extraction

The extracted features can be categorized into 3 groups:

3.3.1. Cyber features:

- Hashtags count: is the count of hashtags per day.
- Favourites count: means how many people favourite the tweets of a day.
- Media count: An array of media is attached to each tweet when downloading it via Twitter search API. For now, only photo type is supported so the media count is the number of photos in a tweet. This feature is aggregated per day by summing up the number of photos in the tweets.
- Source: each tweet has a source; twitter iOS, twitter website, twitter iPad, Instagram, Foursquare, etc. We counted, per day, the number of tweets generated from other application than Twitter.
- Week: it equals 1 if the day is weekday (Monday to Friday) and 0 if weekend (Saturday and Sunday).
- Day_night: it is a ratio of number of tweets at daytime over nighttime. If a tweet is posted between 8 pm and 6 am, it will be considered as nighttime tweet. Otherwise, it is counted at daytime tweet.
- Active_passive: active actions are posts done by user. However, passive actions are retweets.

This feature is a ratio of active actions over the sum of active and passive actions.

3.3.2. Social features:

- Mentions count: it is the number of mentions in the tweets of a day.
- Mention: couple all people mentioned on a day with the corresponding mood.

3.3.3. Physical features:

- Coordinates: we couple the location with mood when a tweet has a geotagged information. It can come from any application (Instagram, Foursquare, Fitbit, etc.) that a user authorizes it to share its location.

All the above mentioned features were normalized with z-score.

4. Results and Discussion

4.1. Correlation

For each user from our dataset, we calculated the Pearson's correlation ($r \in [-1,1]$) between each feature and the mood of next day.

We reported in Table 1 the results with statistically significant correlations of $p < 0.05$ and r above 0.3. Among 800 users, we found 328 users who have at least one feature with correlation above 0.3 ($p < 0.05$). Then, for each feature, we calculate the mean and variance of r.

Results of r prove the existence of significant correlation between the mood and the selected features.

Feature Type	Feature Name	Mean, Variance(r)	Number of users with r>0.3, p<0.05
Cyber	Hashtags count	0.57 , 0.06	35
	Favourites count	0.57 , 0.08	20
	Media count	0.48 , 0.05	32
	Source	0.51 , 0.05	44
	Week	0.50 , 0.04	29
	Day_night	0.52 , 0.06	37
	Active_passive	0.47 , 0.04	36
Social	Mentions count	0.49 , 0.05	50
	Mention	0.48 , 0.02	67
Physical	Coordinates	0.57 , 0.04	104

Table 1: Statistics for Pearson's correlation per feature

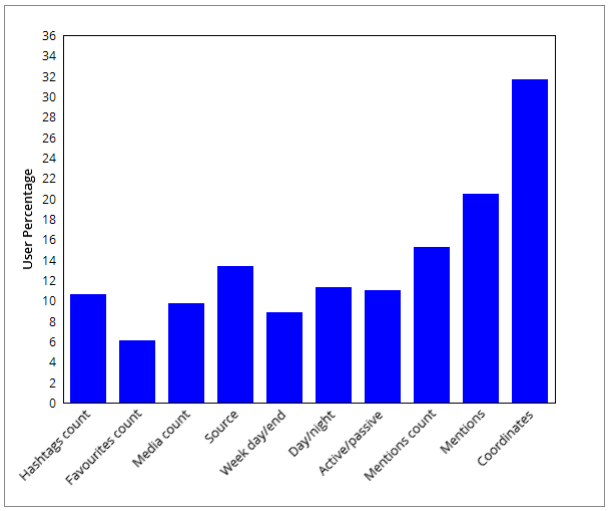
To evaluate the importance of each type of feature, we plot in Figure 2 the percentage of users having significant correlations ($p < 0.05$) and r greater than 0.3 (Figure 2 (a)), 0.5 (Figure 2 (b)) and 0.7 (Figure 2 (c)) for each feature. Results show that the features that appear with a highest percentage, are the one in the social and physical spaces. This result proves the relevance of our claim, that the user mood can be inferred from his physical and social environment when there is lack of activities on cyber space.

4.2. Prediction

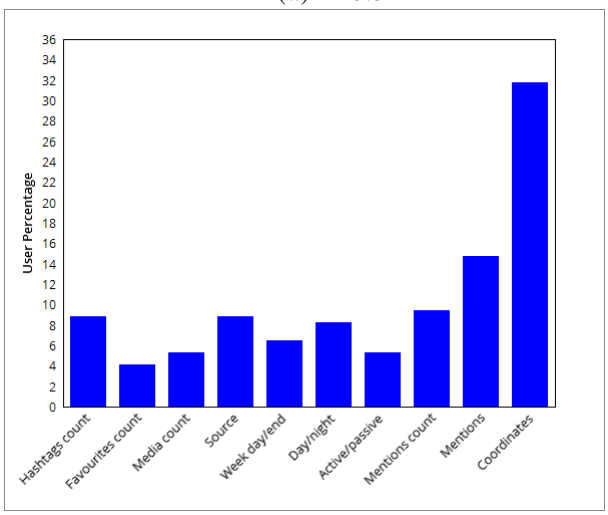
Given that we found significant correlations between cyber, social and physical features and mood of the upcoming day, we trained a predictive system using these features. To evaluate our predictive system, we tried six regression algorithms using Weka, with 10-fold cross-validation. The result is summarized in Table 2. We averaged the root-mean-square error (RMSE) of each classifier over the number of users we have in the dataset. From these results, we can say that our proposed features could successfully predict user upcoming mood.

	Mean (RMSE)	Variance (RMSE)
Linear Regression	0.20	0.018
k-NN: IBk	0.23	0.015
SVR: SMOreg	0.20	0.02
Multi-layer Perception	0.33	0.02
REPTree	0.18	0.008
Random Forest	0.20	0.01

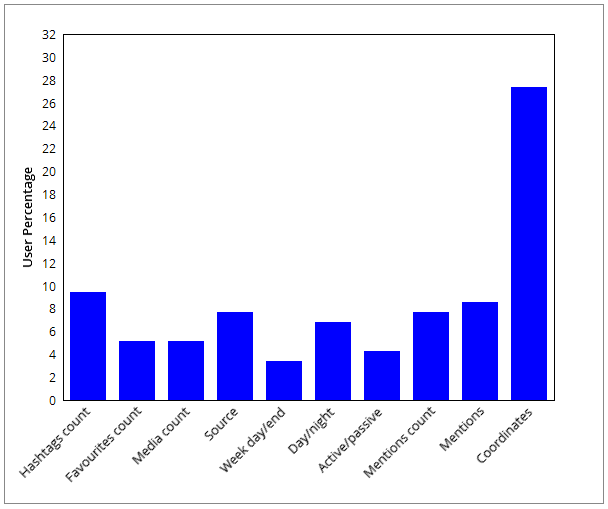
Table 2: RMSE mean and variance for different regression algorithms



(a) r>0.3



(b) r>0.5



(c) r>0.7

Figure 2: Percentage of users per feature with high correlation (r = 0.3, 0.5, 0.7)

5. Conclusion

In this paper, we have shown that a user's mood can be inferred from his cyber, social and physical information he shares on OSNs. With these features, we can predict the mood of the coming day of a specific user even if he wasn't posting text tweets. Our predictive system results in a low RMSE averaged over the number of users studied in our dataset.

However, as mentioned in Section 4.1, 472 users don't have high correlation in our experiment. This is because, in our study, we have just focused on Twitter posts. From this, results were restricted to users sharing their Instagram and Foursquare posts on Twitter. Statistics on Twitter showed that only 20% of online users share their check-ins on Twitter. To increase the number of users with high correlation, we need to collect user posts on his 3 accounts namely, Twitter, Instagram and Foursquare. Expanding the analysis to multiple OSNs is technically possible. That is why, as future work, we want to expand our experiments by analyzing data of each user gathered from multiple OSNs (Foursquare, Instagram, Fitbit and Twitter) accounts.

References

[1] D. Quercia, M. Kosinski, D. Stillwell and J. Crowcroft, "Our Twitter Profiles, Our Selves: Predicting Personality with Twitter", Proc. of the Third International Conference on Social Computing (SocialCom) and the Third International Conference on Privacy, Security, Risk and Trust (PASSAT), IEEE, pp. 180—185, 2011.

[2] B. Ferwerda, M. Schedl, and M. Tkalcic, "Predicting Personality Traits with Instagram Pictures", Proc. of ACM the 3rd Workshop on Emotions and Personality in Personalized Systems (EMPIRE), pp. 7-10, 2015.

[3] M. Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting Depression via Social Media", ICWSM, The AAAI Press, 2013.

[4] M. Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting postpartum changes in emotion and behavior via social media", Proc. of ACM the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 3267-3276, 2013.

[5] Le T. Nguyen, P. Wu, W. Chan, W. Peng, and Y. Zhang, "Predicting collective sentiment dynamics from time-series social media", Proc. of ACM the First International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM), Article 6, pp. 1-8, 2012.

[6] M. Roshanaei, R. Han, S. Mishra, “Features for mood prediction in social media”, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1580-1581, 2015.

[7] J. Alexander Lee, C. Efstratiou, and Lu Bai, “OSN mood tracking: exploring the use of online social network activity as an indicator of mood changes”, Proc. of ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp), pp. 1171-1179, 2016.

[8] M. H. Veiga and C. Eickhoff. “A Cross-Platform Collection of Social Network Profiles”, Proc. of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR), pp. 665-668, 2016.