Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

Predicting Movie Genre Preferences from Personality and Values of Social Media Users

Md. Saddam Hossain Mukta,[1] **Euna Mehnaz Khan,**[2] **Mohammed Eunus Ali,**[3] **Jalal Mahmud**[4]
[1−3]Department of CSE, Bangladesh University of Engineering & Technology, Dhaka, Bangladesh
[4]IBM Research-Almaden, San Jose, CA, USA
[1]saddam944@gmail.com, [2]eunakhan@gmail.com, [3]eunus@cse.buet.ac.bd, [4]jumahmud@us.ibm.com

Abstract

We propose a novel technique to predict a user's movie genre preference from her psycholinguistic attributes obtained from user social media interactions. In particular, we build machine learning based classification models that take user tweets as input to derive her psychological attributes: personality and value scores, and gives her movie genre preference as output. We train these models using user tweets in Twitter, and her reviews and ratings of movies of different genres in Internet movie database (IMDb). We exploit a key concept of psychology, i.e., an individual's personality and values may influence her choice in performing different actions in real life. We have investigated how personality and values independently and collectively influence a user preference on different movie genres. Our proposed model can be used for recommending movies to social media users.

Introduction

Psychological attributes such as personality (Mukta, Ali, and Mahmud 2016a) and values (Chen et al. 2014) of an individual influence her preference and actions while performing different activities. For example, a psychology study (Weaver, Brosius, and Mundorf 1993) finds that user personality, which describes individual pattern of behavior and thoughts, influences user movie preferences as different movies are capable of providing different ranges of stimuli to the user and these stimuli are directly linked to her personality and psychological state. Similarly, values of an individual, describing self-direction, power, and hedonism attributes of a person, influence user behavior and actions such as reading habits, buying products, etc (Verplanken and Holland 2002). In this paper, we propose a novel technique to predict a user's movie genre preference from her psycholinguistic attributes obtained from the social media interactions of the user in two major social media: Twitter and IMDb [1].

In recent years, user tweets have become a primary source of information to find many interesting insights such as preference (Rahman et al. 2016), personality (Mukta, Ali, and Mahmud 2016a), and values (Mukta, Ali, and Mahmud 2016b). These works analyze user tweets to predict user's different psychological attributes. Recently, a couple of studies have been conducted on IMDb, an online database that shares information regarding movies, television program, genres, cast, and biographies, to identify movie genres from movie synopsis (Ho) and IMDb plot key words (Wortman 2010). In an another study, Dooms et al. (Dooms, De Pessemier, and Martens 2013) find structured tweets (e.g., 'I rate matrix 9/10 http://www.imdb.com/title/tt0133093/ #IMDb') containing movie rating of IMDb. Later they predict movie rating in (Dooms and Martens 2014) from the same dataset, namely *movietweetings*[2].

In this paper, we take a step forward by combining user tweets from Twitter and user movie rating in IMDb to build a model that can predict movie genre from user personality and values derived from her psycholinguistic attributes. The key intuition of our approach is that an individual's psychological attributes such as personality and/or values influence her behavior in performing both social and real life activities. To the best of our knowledge, this paper presents the first approach that exploits the fusion of two social media data, i.e., Twitter and IMDb, to predict user movie genre. In particular, we investigate how personality and values independently and collectively influence a user choice for a number of movie genres. A key benefit of our approach is that, by using our proposed model, one can recommend movie to a user by only knowing her personality and/or values from her tweets.

Movies can be classified in different genres based on motion picture category such as action, comedy, science fiction (sci-fi), biographies, and horror. Different people like different genre of movies: some like to watch a movie that makes them think while other people like movies that make them laugh. We observe that IMDb reviews and rating may not be sufficient to predict a user movie preference as these reviews largely represent users' observation regarding the context and plot of those movies. Thus, these reviews do not focus on users' daily preference, psychological states, affective and cognitive mechanisms. However, a recent study (Hsieh et al. 2014) shows that one can accurately identify users' reading interest from psycholinguistic attributes obtained from tweets. Motivated by this study, in this paper, we model the correlation between tweets and movie genre preference

[1]http://www.imdb.com/

[2]https://github.com/sidooms/movietweetings

that can accurately predict the user preference of movie genre from her tweets.

In our study, we have collected data of 232 users who have both Twitter and IMDb profiles. We have a total of 3,65,400 tweets to compute users' personality and value scores. We have collected ratings of 6725 distinct movies of 5 different movie genres. We first compute Big5 personality (John, Naumann, and Soto 2008) traits and Schwartz values (Schwartz et al. 2001) dimensions of users by using IBM Watson personality insights API [3]. Then, we select personality and value features that are highly correlated with users' movie genre preference. After that we build a classification model to predict movie genre preferences from users' personality, and values independently. Finally, we build an ensemble model that combines both personality and values of a user to predict her preference in choosing different movies based on their genre. Personality and values based classifiers have on average AUC scores of 59.4%, and 58.8%, respectively. Ensemble based classifiers have on average AUC score of 63.4%. These AUC scores are significantly better than the baselines.

In summary, we have the following contributions:

- We are the first to exploit the data fusion of Twitter and IMDb to predict users' movie genre preference from their tweets.
- We integrate two psychological attributes: personality and values to build a classification model to predict users' movie preference from user personality and values.
- Our experimental study on 232 users shows the efficacy of our approach.

Methodology

- *Computing personality and values from tweets:* We compute users' personality and value scores by using *IBM Watson personality insights API*.
- *Model building from personality, values, and movie genre:* We select personality and value features that might be correlated with users' movie genre preference. Then we build classification model to predict movie genre preferences from users' personality and values independently.
- *Ensemble of models:* We build an weighted ensemble based predictive model to find users' movie genre preference from personality and values combinedly to improve the accuracy of the prediction.

Data Collection

We first link profiles of IMDb and Twitter users. We extract a total of 251 IMDb users who has Twitter profile under the same *username*. Then we fetch users' movie genres and ratings by using Python *BeautifulSoup* [4] implementation package from IMDb, and collect tweets by using python *tweepy* implementation package. We consider movies that are released in the year of 2015 and 2016, so that reviewers information (i.e., current location) are likely to be updated. We find a total of 13371 different movie reviews. Users rate a total of 6725 distinct movies from their account creation date. There are a total of 27 different genres of movies. Among these 27 genre of movies, we consider only 5 genre of movies as these are most prominent in the dataset. The genres' names and percentages are: *drama (17.90%), thriller (11.53%), comedy (9.14%), action (9.12%), and adventure (7.39%)*, respectively.

Among the collected 251 users, we discard 19 users as they have few tweets to analyze personality and values by using IBM Watson personality insights API. We compute personality and value scores for the remaining of the 232 users. From these 232 users, we find a total of 3,65,400 tweets. The users have maximum, minimum, and average of 3217, 175, and 1575 tweets, respectively.

Feature Selection

Users' movie genre preference may depend on personality, values, or both. According to Big5 model (John, Naumann, and Soto 2008), personality has five different traits: *openness-to-experience, conscientiousness, extraversion, agreeableness, and neuroticism*. Values (Schwartz et al. 2001) have also five broader dimensions: *openness-to-change, hedonism, self-transcendence, conservation, and self-enhancement*. One may think that personality traits may influence a person to choose a movie of a specific genre, while others may think that value dimension may persuade a person to select a certain movie genre. It is difficult to state which traits of personality and values influence a person to watch which genre of movies. Thus, we first need to identify features from both personality traits and value dimensions that are suitable to predict movie genre preference.

We use genre of movies as ground truth data (dependent variable), and personality and value scores as independent variables. We select the best subset of personality and value traits (predictors) using *forward selection* approach of *leaps* [5] R package implementation. Since the forward selection approach starts with no predictors, we can identify which predictors are more significant than the other when these are added stepwise to the list. Leaps package performs an exhaustive search to find out best subset of personality and value dimensions by using an efficient branch-and-bound algorithm. We first select the best subset of 3 features from personality (value dimensions) from the 5 broad personality (values) dimensions independently, and then we identify the best subset of 5 features among the 10 features of personality and values. For example, we select *extraversion, conscientiousness, and neuroticism* personality traits and *hedonism, conservation, and openness-to-change* values independently as features for computing movie genre preference by using the the best subset selection approach. Then we identify the best subset of features for computing movie genre preferences from both personality and values are: *neuroticism, extraversion, agreeableness, hedonism, and self-transcendence.*

[3] https://personality-insights-livedemo.mybluemix.net/

[4] https://pypi.python.org/pypi/beautifulsoup4

[5] https://cran.r-project.org/web/packages/leaps/

Table 1: Personality based classification of movie genres

Genres	Best classifier	AUC	TPR	TNR
Drama	RepTree	0.62	0.94	0.85
Thriller	RepTree	0.55	0.09	0.03
Comedy	RandomTree	0.59	0.06	0.03
Action	RandomTree	0.61	0.28	0.25
Adventure	RepTree	0.60	0.07	0.03

Table 2: Value based classification of movie genres

Genres	Best classifier	AUC	TPR	TNR
Drama	Random Tree	0.63	0.91	0.81
Thriller	Random Forest	0.59	0.17	0.13
Comedy	RepTree	0.59	0.10	0.02
Action	RandomTree	0.54	0.12	0.07
Adventure	Random Forest	0.59	0.07	0.04

Genre Classification

In this section, we first build classification models from personality traits and value dimensions independently. Then, we investigate prediction potential of these classification models by using machine learning techniques. Finally, we build an weighted ensemble of classification model by combining both classifiers of personality and value.

We observe that each movie may fall in the intersection of more than one genre names. For example, *Gone Girl* movie has three different genres: *crime, drama, and thriller*. Since we build classifier for movie genre preference, we need only one class label for each instance. Therefore we distribute the genre names for a single movie name into multiple rows. In this way, we have a total of 39,119 different movie genre instances. In IMDb, users may rate a movie on a scale of 1 to 10. Since we identify users' movie genre preferences, we discard instances that have low rating, i.e., rating <7. When a user gives a low score to a movie genre, it signifies that she has less interest on that movie genre. Finally we have a total 13,264 instances that have a single movie genre name.

Genre classification using personality

We apply Naive Bayes, Support Vector Machine (SVM), Random Forest, Random Tree and RepTree classifiers in our dataset by using WEKA (Hall et al. 2009) machine learning toolkit. Table 1 presents the best classifier, its true positive rate (TPR), true negative rate (TNR) and area under the ROC curve (AUC) for predicting different genre of movies from users' personality traits. TPR defines how many samples are correctly classified as positive among all positive samples and TNR defines how many samples are incorrectly classified as negative among all negative samples available during the test. We conduct the performance of the classifiers by using AUC values under the 10-fold cross validation. We find that on an average the AUC of our classifier is 59.4%. We use *ZeroR* classifier as baseline method, which has an AUC score of 52.9%. We observe that our classifiers always outperform the baseline. We also find that mean absolute error (MAE) of our classification model is 0.29.

Genre classification using values

In this subsection, we again apply Naive Bayes, SVM, Random Forest, Random Tree and RepTree classifiers in our dataset to predict different genre of movies from users' value dimensions. We conduct the performance of the classifiers by using AUC values under the 10-fold cross validation. Table 2 presents best classifier, its TPR, TNR, and AUC for predicting different genre of movies from users' value dimensions. We find that the average AUC of our classifier is 58.8%. The AUC of our baseline is 51.7%. We observe that our classifiers largely perform better than the baseline. We also find that MAE of our classification model is 0.31.

Table 3: Weights (AUC of best classifiers) derived from the weighted training dataset.

Psychological attributes	AUC of best classifier				
	Action	Adven.	Thril.	Dram.	Com.
Personality	0.62	0.63	0.54	0.63	0.55
Values	0.56	0.61	0.61	0.62	0.56

Genre classification using personality and values

We find associations in users' movie genre preferences with their personality traits and value dimensions (according to Section: Feature Selection). Users' different psychological attributes (i.e., personality and values) may contribute differently while identifying movie genre preferences. Among our built classifiers, one classifier may predict a movie genre better than the other. For example, *action* genre of movies can be predicted better with personality traits (61%) than value dimensions (54%); on the other hand, *thriller* genre of movies can be predicted better with value dimensions (59%) than personality traits (55%). Since every psychological attribute contributes to identify users' movie genre preferences based on their strength, we combine all the classification models obtained from the previous independent personality and value models.

Thus, it is necessary to prioritize the psychological attributes based on their importance, as we compute movie genre preferences from personality and values. Performance of individual classifier represents weights of different psychological attributes. To combine personality and value dimensions, we perform the following two steps: i) computing weights for personality traits and value dimensions, and ii) combining the models with an weighted linear ensemble technique.

Learning weights: We determine the weight of personality and value scores to determine preference of movie genres. To this end, we build classification models by using the data of 3980 (30% of the total dataset) movie genre preference instances. To compute weights, we take both personality and value scores as input and the best AUC score of movie genre classifier as output. Table 3 presents the best AUC scores that are computed with different classifiers to predict genre of movie preferences by using 30% of the dataset.

An weighted linear ensemble: We build an weighted linear ensemble model from personality and values of 9284 (70% of the total dataset) instances. We have built dif-

Table 4: Classification result to identify genre by using both personality and value scores

Genres	Best classifier	AUC	TPR	TNR
Drama	RepTree	0.65	0.94	0.85
Thriller	RandomForest	0.59	0.22	0.18
Comedy	RandomTree	0.62	0.27	0.21
Action	RandomTree	0.65	0.52	0.37
Adventure	RepTree	0.64	0.37	0.12

ferent classifiers from personality and values to predict users' movie genre preferences that are described in previous sections. Since we train different classifiers that produce weights, we compute weighted linear ensemble score using the weights in Table 3. Finally, we build our weighted linear ensemble model using the weights generated from another dataset, so that our models do not get over-fitted. Table 4 presents the movie genre preference classification result by using the ensemble of personality and values. We observe that the average AUC of our classifier is 63.4%, and the baseline accuracy is 56.2%. We also observe that our classifiers largely outperform the random baseline. We also find that MAE of our model is 0.18. We find that our ensemble of classifiers achieves higher accuracy than the independent personality and value based classifiers.

Discussion

Our work is the first study to predict movie genre preference from psycholinguistic attributes, i.e., personality and values. We observe in Table 1 that personality trait based genre prediction classifiers perform better than baseline. We see that drama genre classifier shows the strongest (62%) and thriller genre classifier shows the weakest (55%) AUC scores. We also find that classifiers for thriller, comedy, and adventure movie genres show low TPR and TNR scores. On the other hand, we observe in Table 2 that value based classifiers show identical performance to personality trait based classifiers for comedy genre of movies. However, value based classifier shows significant better result for thriller genre of movies and significant weaker result for action genre of movies than that of personality trait based classifier. We find that drama genre classifier shows the strongest (63%) and action genre classifier shows the weakest (54%) AUC scores. Value based classifiers also largely show poor TPR and TNR scores. We observe in Table 4 that personality and values based combined model shows better AUC result than that of independent personality and values based models. These classifiers also show substantial improvement for TPR and TNR scores for movie genre classifiers than personality and values based independeny models.

Conclusion

In this paper, we have predicted movie genre preferences from users' personality and values derived from their social media interactions. We have exploited the data fusion of Twitter and IMDb to build a model that can accurately predict movie genre preferences from user personality and values. We have demonstrated which types of personality traits and value dimensions better predict which type of movie genre by using classification techniques. In future, we plan to integrate our model with a real movie recommendation application for social media users.

Acknowledgment

This research is funded by ICT Division, Ministry of Posts, Telecommunications and Information Technology, Government of the People's Republic of Bangladesh.

References

Chen, J.; Hsieh, G.; Mahmud, J. U.; and Nichols, J. 2014. Understanding individuals' personal values from social media word use. In *CSCW*, 405–414.

Dooms, S., and Martens, L. 2014. Harvesting movie ratings from structured data in social media. *ACM SIGWEB Newsletter* (Winter):4.

Dooms, S.; De Pessemier, T.; and Martens, L. 2013. Movietweetings: a movie rating dataset collected from twitter. In *Workshop on CrowdRec at RecSys*, volume 2013, 43.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsl.* 11(1):10–18.

Ho, K.-W. Movies genres classification by synopsis.

Hsieh, G.; Chen, J.; Mahmud, J. U.; and Nichols, J. 2014. You read what you value: understanding personal values and reading interests. In *CHI*, 983–986.

John, O. P.; Naumann, L. P.; and Soto, C. J. 2008. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research* 3:114–158.

Mukta, M. S. H.; Ali, M. E.; and Mahmud, J. 2016a. Identifying and validating personality traits-based homophilies for an egocentric network. *Social Network Analysis and Mining* 6(1):74.

Mukta, M. S. H.; Ali, M. E.; and Mahmud, J. 2016b. User generated vs. supported contents: Which one can better predict basic human values? In *SocInfo*, 454–470. Springer.

Rahman, M. M.; Majumder, M. T. H.; Mukta, M. S. H.; Ali, M. E.; and Mahmud, J. 2016. Can we predict eat-out preference of a person from tweets? In *Websci*, 350–351. ACM.

Schwartz, S. H.; Melech, G.; Lehmann, A.; Burgess, S.; Harris, M.; and Owens, V. 2001. Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of cross-cultural psychology* 32(5):519–542.

Verplanken, B., and Holland, R. W. 2002. Motivated decision making: effects of activation and self-centrality of values on choices and behavior. *Journal of personality and social psychology* 82(3):434.

Weaver, J. B.; Brosius, H.-B.; and Mundorf, N. 1993. Personality and movie preferences: A comparison of american and german audiences. *Personality and Individual Differences* 14(2):307–315.

Wortman, J. 2010. *Film classification using subtitles and automatically generated language factors*. Technion-Israel Institute of Technology, Faculty of Industrial and Management Engineering.