Empathic inclination from digital footprints*

Marco Polignano, Pierpaolo Basile, Gaetano Rossiello, Marco de Gemmis, and Giovanni Semeraro

University of Bari "Aldo Moro", Dept. of Computer Science
name.surname@uniba.it

Abstract. *The large amount of personal data left by users on the Internet is a valuable source of information for improving the efficacy of profiling tasks. In particular, the data collected from social media can disclose personal habits, preferences and affective traits. The study is focused on the emphatic inclination of a subject, i.e. the ability to feel and share another person's emotions, which can be a relevant aspect to consider in retrieval or recommendation processes. To support this idea, a model was proposed to predict its level and to emphasize the correlations with explicit features that characterize the user.

Keywords: Social medium footprint, Empathy, Machine Learning

1 Background and Motivations

The massive spread of social media over mobile devices has significantly changed the way people communicate today. The interaction with social media allows a person's to feed her digital identity with preferences, interests, aptitudes. That information, usually known as social media footprints, is available on the web and can be exploited by others to discover that person's tendencies, styles of life, and also affective and psychological traits [3, 7]. For this reason, we want to investigate whether (and how) it is possible to predict the empathy inclination of a user. We believe that personalization systems working in some specific domains, such as movie or music recommendation, would benefit from the knowledge of this affective aspect of the user. According to Hogan [2], empathy can be correlated with social self-confidence, even-temperedness, sensitivity and nonconformity. Therefore, a subject who shows high empathy is a very emotional and sensitivity person because not only she is inclined to understand others' emotions, but she is also able to feel some strong emotions for them.

2 Empathy Inclination Prediction Model

The proposed model is based on the idea that several aspects of the user life might contribute to infer user inclination to empathy. We exploit several kinds

* These results are already published in "Learning Inclination to Empathy from Social Media Footprints" in proceedings of User Modelling, Adaptation and Personalization, FIIT STU, Bratislava, Slovakia, July 2017 (UMAP 2017)

of features, as sketched in Fig.1, to predict an empathy score by different linear regression models.

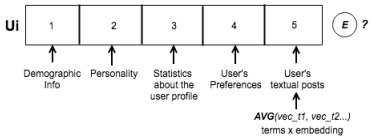

Fig. 1. Empathy prediction model

Each user U_i is represented as the concatenation of five features vectors. Each vector captures a particular aspect of the user profile. User's preferences are obtained by analyzing her likes grouped by topics over social media and they include likes over pages, artists, movies and many other topics of interest. The representation used is the SVD [5] for the representation through relevant combinations of *concepts* and LDA [1] for a combination of *descriptive topics*. The posts are analyzed by a pipeline performing basic NLP operations (we adopted TweetNLP as tokenizer: `http://www.cs.cmu.edu/~ark/TweetNLP/`), as well as operations for annotating emoticon and for removing character repetitions longer than two inside words. In order to capture the semantics behind the words, we use the word2vec algorithm [6] over all the textual posts in the collection for learning 200-dimension vectors, by considering only words that occur at least 10 times and 10 epochs of learning. Moreover, we divide the whole vocabulary of word2vec vectors into clusters, which should represent topics of discussion.

3 Experimental Session

The aim of the experiment is to predict the user's empathy by exploiting information explicitly available on her Facebook profile, as well as implicit information that can be inferred, as explained in Sec. 2. Moreover, we want to identify which groups of features are more important for obtaining an accurate prediction, by discovering relevant correlations among empathy and user's features.

More precisely, we formulated the following research questions:

- **RQ1.** Is it possible to predict empathy from social media footprints?
- **RQ2.** What are the most important features to consider for improving the prediction accuracy?

The dataset used in the experiment, proposed by Kosinski [4], contains information about 4 millions of Facebook users. Data are collected using the *"myPersonality"* Facebook application. We removed those users who have not terminated the questionnaire or who were not linkable to other data (Demographic, Personality Traits, Activity, Status), after this step, the dataset is composed by 903 users, 178,766 status updates. The range of the empathy value is 0-80. We

exploit three different regression algorithms: 1) *Linear Regression (Lr)*, 2) *Simple Regression (Sr)* , 3) different configurations of kernel of the *SVM Regression with SMO algorithm (SMO)* . For the SMO we used the polynomial kernel (SMO_{poly}) and the Radial Basis Function (RBF) kernel (SMO_{rbf}), by varying the c parameter from 1 to 8. We propose two simple baselines in order to compare the proposed approach with alternative options. The former always predicts the most frequent value in the dataset *(Majority, Value Predicted= 8, MAE= 7.4784, RMSE= 10.8258)*, while the latter computes the empathy score as the simple average of EQS observed in the dataset *(Avg EQS, Value Predicted= 13.9169, MAE= 6.8457, RMSE= 9.0757)*. As for the evaluation metrics, we adopted the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE).The evaluation protocol was 10 folds cross validation.

4 Discussion of Results

We execute a first experiment by running Lr, Sr, and SMO by using all the features of the dataset (1088 features in total). We compared the results in Tab. 1 with our baselines observing that using SMO with a polynomial kernel is not a good choice, having a large number of features. On the contrary, SMO with an RBF kernel is able to overcome both the baselines by setting $c = 1$ ($MAE = 5.9101$, $RMSE = 8.2341$). These results allow us to answer positively to **RQ1**. Interesting results are obtained by Sr. MAE and RMSE are better than the baselines, despite this algorithm creates a regression function considering only the feature with higher variance in the dataset. Due to these findings, we decided to perform feature selection. We exploit the correlation-based feature subset

Table 1. Relevant results of empathy level prediction

		All Features		*Filtered Features*	
Approach	**c**	**MAE**	**RMSE**	**MAE**	**RMSE**
SMO_{poly}	1	12.7137	19.1565	***5.714***	***7.8407***
SMO_{rbf}	2	5.9543	8.2432	***5.6673***	7.8631
SMO_{rbf}	8	6.539	8.7748	5.686	***7.8236***
Lr	-	22.7929	34.4679	5.7854	7.7269
Sr	-	6.1045	8.233	6.1045	8.233

selection for finding the set of "most informative" features for the prediction task. The selected features are those with high correlation with the prediction class and low correlation among them. We obtained a set of 37 features. The best result in term of MAE (5.6673) is obtained by the SMO_{rbf}, with $c = 2$. This configuration does not provide the best RMSE (7.8236) that it is achieved by SMO_{rbf} with $c = 8$. For the SMO_{poly} configuration, the best result for both MAE and RMSE is obtained with $c = 1$ (5.714, 7.8407). It is interesting to note that results obtained by exploiting only selected features are better than both the baselines

and the runs over the whole set of features. Analyzing the features emerged after the selection process, we can note some interesting correlations among the semantics of them and the empathy inclination of the user. In particular, we observed that for an accurate prediction we have to consider the user's *religion* (Nonreligious/Atheist), *country* (AG, EG, KW, HN, AR, SR), *relationship_status* (Separated), *personality* (extroversion, agreeableness) and some relevant word2vec clusters: *cluster_1*: game, team, soccer, battle, race, fans, bowling; cluster_13: dear, cheers, goody, extraordinaire, excitedly; *cluster_21*: personality, motivation, destiny, ability, vision; *cluster_24*: facebook, phone, message, internet, video. These correlations can be used as hints for user profiling and *partially* provide an answer for **RQ2**, therefore we decided to perform an ablation analysis for further investigation. We selected the best configuration SMO_{rbf} with $c = 1$ and we removed one set of features at a time. By removing groups of features such as *demographic, activity, LDA*, we observed a slight change of MAE and RMSE. On the contrary, by removing the set of features about *personality*, a significant increase of both MAE (9.6308) and RMSE (9.0815) is observed. This provides a more specific answer for **RQ2**: **personality traits** are the key for effective empathy prediction.

5 Conclusion

In this paper, we investigated the problem of mining social media footprints to infer the user's inclination toward empathy. The main outcome of the experiments is a strong correlation is observed among empathy and personality traits. As a future work, we plan to include the findings described in this preliminary study as part of the user profile and to include them in a recommendation strategy.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)
2. Hogan, R.: Development of an empathy scale. Journal of consulting and clinical psychology 33(3), 307 (1969)
3. Jelenchick, L.A., Eickhoff, J.C., Moreno, M.A.: Facebook depression? social networking site use and depression in older adolescents. Journal of Adolescent Health 52(1), 128–130 (2013)
4. Kosinski, M., Matz, S.C., Gosling, S.D., Popov, V., Stillwell, D.: Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. American Psychologist 70(6), 543 (2015)
5. Landauer, T.K.: Latent semantic analysis. Wiley Online Library (2006)
6. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
7. Skowron, M., Tkalčič, M., Ferwerda, B., Schedl, M.: Fusing social media cues: personality prediction from twitter and instagram. In: Proceedings of the 25th international conference companion on world wide web. pp. 107–108. International World Wide Web Conferences Steering Committee (2016)