Automatic Estimation of Web Bloggers' Age Using Regression Models

Vasiliki Simaki(✉), Christina Aravantinou, Iosif Mporas, and Vasileios Megalooikonomou

Multidimensional Data Analysis and Knowledge Management Laboratory
Department of Computer Engineering and Informatics,
University of Patras, 26500 Rion, Greece
{simaki,aravantino,vasilis}@ceid.upatras.gr, imporas@upatras.gr

Abstract. In this article, we address the problem of automatic age estimation of web users based on their posts. Most studies on age identification treat the issue as a classification problem. Instead of following an age category classification approach, we investigate the appropriateness of several regression algorithms on the task of age estimation. We evaluate a number of well-known and widely used machine learning algorithms for numerical estimation, in order to examine their appropriateness on this task. We used a set of 42 text features. The experimental results showed that the Bagging algorithm with RepTree base learner offered the best performance, achieving estimation of web users' age with mean absolute error equal to 5.44, while the root mean squared error is approximately 7.14.

Keywords: Author's age estimation · Text processing · Regression algorithms

1 Introduction

The extensive growth of the web and the plethora of options that social media provide, have resulted in the increase of the web users population, especially in the most developed countries. This reality results to the production of large amounts of written web posts on a daily basis. The automatic extraction of information from these online data is related not only to the text itself but also to the gender, age and other demographic characteristics of the user that are essential in the e-government, security and e-commerce market.

The detection of demographic information and more specifically the detection of age, among social media users may be important not only for commercial and sociological purposes, but also for security reasons. Teen users are allowed to use social media without often being supervised by adults, a situation that can be fatal in extreme conditions. It is thus important to be able to automatically estimate the age of an internet user from his/her writing input on the web. Except security, the estimation of the user's age can be important in detecting

© Springer International Publishing Switzerland 2015
A. Ronzhin et al. (Eds.): SPECOM 2015, LNAI 9319, pp. 113–120, 2015.
DOI: 10.1007/978-3-319-23132-7_14

the different trends, opinions, political and social views of each age group. This can enable social scientists to derive important clues about the anthropography among social media users, and how different age groups behave online. Market analysts and advertisers may also be interested in this kind of studies, in order to promote their product or a service in an age-targeted way according to their expressed interests and opinions.

Most studies on age identification treat the issue as a classification problem. In this article, instead of following an age category classification approach, we investigate the appropriateness of several regression algorithms on the task of age estimation of bloggers, dealing with a numerical estimation problem. We relied on several text-based features that have been widely used in the literature for text classification, authorship attribution, gender and age identification, in order to evaluate the performance of regression methods. The remainder of this paper is organized as follows: Sect. 2 presents the state-of-the-art in theoretical and automatic age estimation. Section 3 describes the followed methodology for age estimation from web posts. Section 4 presents the experimental setup and the achieved results. Finally Sect. 5 concludes this work.

2 Background Work

People of different age, gender, educational level, professional activity and geographical orientation make various linguistic choices, due to these social factors [1]. The matching of a linguistic attitude to the corresponding social group is one of the objectives of sociolinguistics. Several sociolinguistic studies in age variation [2,3] observed that teenagers use the language in a more creative and non-contractual way, by producing new forms, when adults prefer more standard types. Semantic neologisms, slang types, loanwords and code expressions are produced by teens, when adults tend to have a more conservative linguistic attitude. This can be explained after the social role in the production/work cycle and the family responsibilities that adulthood occurs, when teens and older people let to a more "loose" use of language [4].

Whilst sociolinguistic researches in age variation stand on theoretical and empirical findings, recent studies in text mining use machine learning algorithms and natural language processing methods for the automatic estimation of the authors' age. Schler et al. [5] create the "Blog Authorship Corpus" in order to identify the author's age and gender. They used style-related features and content-based characteristics in order to detect the gender and the age. They observed that specific forms and unigrams are more frequent in young bloggers, the blogging style and topics are different among 10's, 20's and 30's. Argamon et al. [6] used the corpus from their previous study [5], in order to go deeper in the gender and age mining from text. They used stylistic and content-based features in order to demonstrate the significant variation between different genders and ages in blogging. Goswami et al. [7] performed a stylometric analysis in terms of gender and age by using non-dictionary forms and the sentence length as features. The slang, smileys, out-of-dictionary words, chat abbreviations, on the one hand, and

the sentence length on the other, proved to be highly distinctive among different ages and gender. Tam and Martell [8] performed age classification experiments, using Bayesian and SVM classifiers. They extracted character n-grams and word meta-data features, in order to classify the "NPS Chat Corpus" into five age groups. In their work, Peersman et al. [9], implemented age classification in small texts, using chat words as features, along with character-based features, achieving more than 88 % of accuracy. Other studies in age prediction [10,11], proved that content and stylistic features are extremely significant, and when the online users' activity is added, the classification accuracy increases approximately to 80 % [10]. In their overview of PAN 2013, Rangel et al. [12] presented the different feature sets that the participants in the Author Profiling Task used, which were finally grouped into stylistic-based, content-based, n-grams-, IR-, and collocations-based. Many of the participants dealt with the age detection, and Flekova & Gurevych [13] focused on age and gender using surface, syntactic and punctuation, readability, semantic, content, lexical and stop words features. They observed eventually that the age and gender profiling are not independent issues, but they are determined by the same features. Rangel & Rosso [14] used the PAN-AP-13 dataset in order to perform classification experiments in terms of age and gender, using though features based in cognitive traits of neurology studies. Their approach was more efficient in age than gender prediction and they proved the differences in language use of different ages, in English and Spanish. [15] is a quite integrated study in personality, gender and age detection of Facebook users. Standard approaches were implemented and a particular method was proposed for linguistic analysis and evaluation in terms of personality, age and gender with reliable results, contribution to interdisciplinary researches, and suggestion of new hypotheses and insights. Nguyen et al. [16] performed a study in language use among different age categories of Twitter users. Their analysis showed that differences in style, references, conversation and sharing depended not only on the age category estimation, but also on the life stage and the actual age of the user. Lately, the authorship profiling has become a task about multilingual efforts and [17] is one of the several studies implemented in a non-English corpus for stylometric research and possibilities to perform age, gender, opinion, authorship and personality experiments.

3 Proposed Age Estimation of Web Bloggers Using Regression Models

The estimation of the age of an author is a numerical estimation problem. Although some of the related work found in the literature targets at identifying the age class the author belongs to, according to some quantization of the age scale to age intervals of interest, we target at the direct estimation of the age value of the web blogger (i.e. the author). Thus the problem is formulated as follows: We consider a representation of each web blog post with a feature vector V_n, for the n^{th} post with $1 \leq n \leq N$. A machine learning regression algorithm, f, is used as a numerical estimator, for assigning an age estimation, u, to each feature vector V_n, i.e. $u = f(V_n)$.

For the representation of each web blog post with a feature vector, we used a number of well-known and widely used in text-based analyses features, which have been used in the tasks of author, gender and age identification, they are normalized and they are presented in Table 1. The resulted feature vector has length equal to 42, i.e. $V_n \in \Re^{42}$.

Table 1. The description of features used in our study.

# of characters	# of the "hapax legomena"
# of alphabetic characters	# of the "hapax dislegomena"
# of upper case characters	# of pronouns
# of digit characters	# of function words
# of space characters	average # of sentences per paragraph
# of tab ("\t") characters	average # of characters per paragraph
# of occurrence of each alphabetic character	# of words starting with a capital
# of occurrence of special characters	# of punctuation symbols
# of words	# of capitalized types
# of words with length less than 4 characters	std of the word length
# of characters per word	max length word
average word length	min length word
# of sentences	# of future tense types
# of paragraphs	# of hyperlinks
# of lines	# of self-references
average # of characters per sentence	# of nouns
average # of words per sentence	# of proper nouns
# of unique words	# of adjectives
# of articles	# of prepositions
# of adverbs	# of emoticons
# of interjections	# of verbs

For the regression stage, we relied on a number of dissimilar machine learning algorithms, which have extensively been reported in the literature. In particular, we used:

- The multilayer perceptron neural network (MLP) with three layers which is capable for numerical predictions [18], since neurons are isolated and region approximations can be adjusted independently to each other,
- The support vector machines (SVM) for regression using the sequential minimal optimization algorithm and two different kernels, the radial basis kernel (rbf) and the polynomial kernel (poly),
- The M5 model tree (M5P) algorithm, which is a rational reconstruction of M5 method,
- The K-nearest neighbors algorithm (IBk),
- The RepTree, a fast decision tree learner, which builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with back-fitting). RepTree only sorts values for numeric attributes once

and missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5),
- The Additive regression meta-classifier that enhances the performance of a regression base classifier along with DecisionStump, SVMs with polynomial and radial basis kernel and REPTrees,
- The bagging algorithm combined with REPTree and SVMs with polynomial and radial basis kernel, aiming to reduce variance,
- The M5Rules, which generates a decision list for regression problems using separate-and-conquer. In each iteration, it builds a model tree using M5 and makes the "best" leaf into a rule.

All regression algorithms were implemented using the WEKA machine learning toolkit [19].

4 Experimental Setup and Results

For the present evaluation we used the "Blog Authorship Corpus" [5], a collection of blog posts from 19,320 bloggers which have posted in their blogs. These blog posts were gathered from blogger.com in August 2004. The size of the corpus is 681,288 posts and over 140 million of words, which corresponds to 35 posts and 7,250 words per person. The bloggers fall into three age categories: 10's, 20's and 30's. The 10's age group is constituted of 8,240 blogs whose authors are between 13 and 17 years. The 20's is constituted of 8,086 blogs of 23–27 years old authors. Finally the 30's age group contains 2,994 blogs produced by bloggers between 33 and 47 years. Each blog is structured in a separate file containing the bloggers' posts, the bloggers' id number, his/her gender, his/her exact age and in many cases other anonymised personal pieces of information.

The "Blog Authorship Corpus" was evaluated on the task of age estimation, using the features described in the previous section. The performance of the evaluated regression algorithms was measured using the mean absolute error (MAE) and the root mean squared error (RMSE) of the difference (i.e. the error) between the actual and the estimated age of each web blogger. In order to avoid overlap between training and test subsets, a 10-fold cross validation evaluation protocol was followed. The experimental results for the evaluated regression algorithms in terms of MAE and RMSE are tabulated in Table 2. The best performance for each of the above metrics is indicated in bold.

As can be seen in Table 2, the best performing algorithm was the Bagging implemented with the RepTree base learner, achieving MAE and RMSE equal to 5.44 and 7.15, respectively. The second and third best performance was achieved by the RepTree regression algorithm and the Additive Regression algorithm with the RepTrees regression base classifier with MAE approximately equal to 5.67. The results show the appropriateness of RepTree regression algorithm for the task of age estimation from web blog posts, since it outperforms all the other algorithms either as a base learner within a meta-classification scheme or as a standalone regression algorithm. The superiority of the RepTree regression algorithm is not restricted only in the MAE criterion, but is also presented in

Table 2. Age estimation MAE and RMSE per regression algorithm.

Regression Algorithm	MAE	RMSE
MLP	8.013	10.2701
SVM-poly	5.8917	7.6787
SVM-rbf	6.0666	7.7548
IBk	7.5044	10.2675
Add. Regression (SVM-poly)	5.8702	7.6929
Add. Regression (SVM-rbf)	5.6772	8.0213
Add.Regression (RepTrees)	5.6741	7.5382
Bagging (SVM-poly)	5.9164	8.053
Bagging (SVM-rbf)	6.7432	8.345
Bagging (RepTrees)	**5.4407**	**7.1457**
M5Rules	24.5921	1196.7574
M5P	18.864	863.185
RepTrees	5.6695	7.464

the RMSE criterion, which shows that RepTrees offer the minimum outliers in terms of age estimation comparing to the rest of the evaluated algorithms.

The only regression algorithm which was found to have performance comparable to RepTrees was the SVM with polynomial kernel, performing slightly worse both as standalone and as base learner of a meta-classification algorithm. The good performance of the SVM algorithm is probably owed to the fact that they don't suffer from the curse of dimensionality. The worse performance was achieved by the M5Rules and M5P regression algorithms, which are model trees in contrast to the best performing RepTree which is a regression tree. Their low performance is probably owed to the fact that they leverage potential linearity at leaf nodes and the fact that they construct hard-decision rules based on the best leaf.

5 Conclusion

We presented an evaluation of regression algorithms for the estimation of web bloggers' age. For the estimation of the age we relied on a number of text-based characteristics, which are typical in text classification tasks related to gender and age identification, and constructed one feature vector for each blogger's post. The evaluation results showed that by using regression methods, age estimation can be adequately performed. The RepTree algorithm proved to outperform all the evaluated regression algorithms, and achieved accurate age estimations both when used as main regression algorithm and as a base learner of a meta-classification method. The application of regression algorithms on age categories dramatically increased age estimation accuracy both in terms of mean absolute error and in terms of root mean square error, which indicates that the combination of age category classification followed by age regression per category would offer robust estimation of the web bloggers' age.

References

1. Labov, W.: Sociolinguistic Patterns (No. 4). University of Pennsylvania Press, Philadelphia (1972)
2. Trudgill, P.: The social differentiation of English in Norwich, vol. 13. CUP Archive, Cambridge (1974)
3. Eckert, P.: Age as a sociolinguistic variable. In: Coulmas, F. (ed.) The Handbook of Sociolinguistics. Blackwell, Oxford (1997)
4. Labov, W.: Principles of linguistic change, cognitive and cultural factors, vol. 3. John Wiley & Sons, New York (2011)
5. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, vol. 6, pp. 199–205 (2006)
6. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Mining the blogosphere: age, gender and the varieties of self-expression. First Monday, 12(9) (2007)
7. Goswami, S., Sarkar, S., Rustagi, M.: Stylometric analysis of bloggers' age and gender. In: Third International AAAI Conference on Weblogs and Social Media (2009)
8. Tam, J., Martell, C.H.: Age detection in chat. In: IEEE International Conference on Semantic Computing, ICSC 2009, pp. 33–39. IEEE (2009)
9. Peersman, C., Daelemans, W., Van Vaerenbergh, L.: Predicting age and gender in online social networks. In: Proceedings of the 3rd international workshop on Search and Mining User-Generated Contents, pp. 37–44. ACM (2011)
10. Rosenthal, S., McKeown, K.: Age prediction in blogs: a study of style, content, and online behavior in pre-and post-social media generations. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 763–772. ACL (2011)
11. Nguyen, D., Smith, N.A., Ros, C.P.: Author age prediction from text using linear regression. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 115–123. ACL (2011)
12. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at PAN 2013. Notebook Papers of CLEF (2013)
13. Flekova, L., Gurevych, I.: Can we hide in the web? Large scale simultaneous age and gender author profiling in social media. In: CLEF 2012 Labs and Work-shop. Notebook Papers (2013)
14. Rangel, F., Rosso, P.: Use of language and author profiling: identification of gender and age. Natural Language Processing and Cognitive Science, 177 (2013)
15. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Ungar, L.H.: Personality, gender, and age in the language of social media: the open-vocabulary approach. PloS one **8**(9), e73791 (2013)
16. Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: "How old do you think i am?"; A study of language and age in twitter. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media. AAAI Press (2013)
17. Verhoeven, B., Daelemans, W.: CLiPSStylometry Investigation (CSI) corpus: a Dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (2014)

18. Chester, D.L.: Why two hidden layers are better than one. In: Proceedings of the International Joint Conference on Neural Networks, vol. 1, pp. 265–268 (1990)
19. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Elsevier, Morgan-Kaufman Series of Data Management Systems, San Francisco (2005)