Employing Relation Between Reading and Writing Skills on Age Based Categorization of Short Estonian Texts

Avar Pentel

Tallinn University, Institute of Informatics, Estonia

pentel@tlu.ee

Abstract. In this paper, we present results of our study on age-based categorization of short texts as 85 words per author. We introduce a novel set of features that will reliably work with short texts, and is easy to extract from the text itself without any outside databases. These features were formerly known as variables in readability formulas. We tested datasets presented two age groups - children and teens up to age 15 and adults 20 years and older. Besides readability features, we also tested widely used n-gram features. Models trained on readability features performed better or as well as models trained on n-gram features. Model generated by Support Vector Machine with readability features yield to f-score 0.953.

Keywords: age detection, readability features, n-grams, logistic regression, support vector machines, bayesian.

1 Introduction

With a wide spread of social media, growing problem is related to false identities. Younger people might pretend adults to access adult sites, and older people might pretend youngsters to communicate with youngsters. As we can imagine, this might lead to serious threats, as for pedophilia or other criminal activities. Thus, automatic age detection has serious practical application in social media.

While many works are published on text authorship profiling, social media poses two problems that are not solved this far.

The first problem is related to the amount of the text needed to make predictions. Usually a large training data sets and long texts per author are used [1,2] to make such classification models, but in social media, we can only relay on short texts.

The second problem is related to the cost of feature extraction. Most of the recent studies [3-6] on age detection using word and character n-gram based features and additional databases or systems, as part of speech tagging, etc., to assess the roles of the words in a sentence. With millions of users, these techniques are too costly to be applicable. Ideally, a system could use mostly client side resources.

In this paper, we propose a novel set of features for author's age based profiling that solves both previously mentioned problems. We call these new features the read-

ability features. These features can be easily extracted using client side JavaScript and they make at least as best classifiers as widely used n-gram based features.

We suppose that authors reading skills and writing skills are correlated, and by analyzing author's text readability, we can conclude about his/her education level, which at least to the particular age is correlated with the actual age of an author. Therefore, we can employ old readability formulas that were developed already before computerized era. Automated Readability Index [8], Gunning Fog index [9], SMOG [10], Flesch-Kincaid [11], and other readability formulas assess how much education is needed to understand particular texts. If we take a closer look at the first pair of these formulas (1,2), we can see, that they are using very simple variables, which can be easily extracted from text.

$$ARI = 4.71 \times \left(\frac{characters}{words}\right) + 0.5 \times \left(\frac{words}{sentences}\right) - 21.43 \quad (1)$$

$$GFI = 0.4 \times \left[\left(\frac{words}{sentences}\right) + 100 \times \left(\frac{complexwords}{words}\right)\right] \quad (2)$$

As readability indexes are developed for texts with about 100 words, these are good candidates for our task.

2 Methodology

We collected short texts, average 85 words long, from different social media sources like Facebook, Blog comments, and Internet forums. All authors were identified, and they used in their texts Estonian language. We chose balanced and stratified dataset with 400 instances and with different age groups: 7-15 and 20-48.

We used three types of features in our training datasets: readability features, character n-grams and word n-grams.

Readability features are quantitative data about texts, as for instance an average number of characters in a word, syllables in word, etc. All together 14 different features were extracted from each text as shown in Table 1.

Table 1. Readability features

feature	explanation	calculation	feature	explanation	calculation
CPW	average number of characters per word	$=\frac{Characters}{Words}$	S1TW	words with 1 syllable to all words ratio	$=\frac{1SylWords}{Words}$
WPS	average number of words per sentence	$=\frac{Words}{Sentences}$	SnTW	words with n (2-8+) syllable to all words ratio	$=\frac{nSylWords}{Words}$
CPS	average number of commas per sentence	$=\frac{Commas}{Sentences}$	CWPS	average number of complex words in sentence	$=\frac{ComplexWords}{Sentences}$
SPW	average number of syllables per word	$=\frac{SyllablesInText}{Words}$	CWTW	complex words to all words ratio	$=\frac{ComplexWords}{Words}$

Complex word in our feature set, is a loan from Gunning Fog Index [9], where it means words with 3 or more syllables. As in the Estonian language average number of syllables per word is higher, we raised the number of syllables accordingly. We also created a new and very simple syllable counter for Estonian language.

Another type of features we used, are character n-grams. We extracted all occurred character bigrams and trigrams and using X^2 attribute evaluation, we selected 119 character bigrams and 576 character trigrams.

Similarly, we extracted all occurred word unigrams, bigrams and trigrams and using X^2 attribute evaluation, we selected as features 100 word unigrams, 30 word bigrams and 6 word trigrams.

We made four different datasets: with readability features, with character n-grams, with word n-grams, and with all features combined. The models were generated using Support Vector Machine, Logistic Regression and Naïve Bayes algorithm. Motivation of using these algorithms comes from the literature [12]. Java implementations of listed algorithms that are available in the Weka [13] library were used. 10-fold cross validation was used for evaluation.

3 Results

As shown in Table 2, readability features trained a better classifier with Support Vector Machines and Logistic Regression, yielding to f-scores 0.953, and 0.95 accordingly. Naïve Bayes performed better with n-gram features. Combined feature sets did not improve the models.

Table 2. Results of models trained with different feature types

Classifier	F-Scores			
	Readability	***Char n-grams***	***Word n-grams***	***All combined***
SVM standardized	0.953	0.952	0.850	0.950
Logistic Regression	0.950	0.929	0.775	0.920
Naïve Bayes	0.811	0.946	0.901	0.882

Most distinctive features, among readability features were average number of words in a sentence and average number of characters in a word.

Using logistic regression model with readability features, we created a prototype client side age detection application [14].

4 Conclusion

Employing relations between reading and writing skills, and using features from old readability formulas proved to be an effective way to predict author age class. Readability features are in many ways favorable. First, they are easy to extract, they are self sufficient, and can be computed without any extra help. Syllable counting is

problematic for some languages, but maybe it can be omitted, as syllable count is also not used in all readability indexes.

Secondly, when dealing with short texts, content-based features, as n-grams tend to be very context dependent, the topic can cause a rise of frequency of some words that can be associated to a particular age group. It seems, that how we write depends less on the context than what we write.

However, we have to address limitations of the current study. First, it is obvious, that we cannot use readability features to categorize older age groups. For most of the people, reading and writing skills will not improve continuously during the whole life.

Secondly, it is possible that good age based categorization results are caused by some specific property of Estonian language. For example, Estonian language has many agglutinative inflectional suffixes, and therefore grammatical richness yield directly to more syllables and longer words. Therefore, we look forward to test how readability features work with other agglutinative and inflectional languages.

References

1. Burrows, J. All the way through: testing for authorship in different frequency strata. Literary and Linguistic Computing. 22, 1, pp. 27–47. Oxford University Press (2007)
2. Sanderson, C., and Guenter, S. Short text authorship attribution via sequence kernels, Markov chains and author unmasking: an investigation. EMNLP'06. Association for Computational Linguistics. pp. 482–491. Stroudsburg, PA, USA (2006)
3. Peersman, C., Daelemans, W., Vaerenbergh, L. Predicting age and gender in online social networks, SMUC '11 Proceedings of the 3rd international workshop on Search and mining user-generated contents, pp. 37-44. (2011)
4. Argamon, S., et al. Automatically profiling the author of an anonymous text. Communicatin of the ACM 52(2) pp. 119–123 (2009)
5. Nguyen, D., Rose, C.P. Age prediction from text using linear regression. LaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp 115-123 (2011)
6. Weren, E.R.D. et al. Using simple sontent features for the author profiling task. Notebook for PAN at CLEF, (2013)
7. Marquart, J. et al. Age and gender identification in social media. CEUR Workshop Proceedings, vol 1180 (2014)
8. Senter, R.J., Smith, E.A. Automated Readability Index. Technical report, Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, Ohio (1967)
9. Gunning, R. The Technique of Clear Writing. New York: McGraw–Hill (1952)
10. McLaughlin, G. Harry. SMOG Grading - a New Readability Formula. Journal of Reading 12 (8): 639–646 (1969)
11. Flesch, R. A new readability yardstick. Journal of Applied Psychology 32: pp. 221–233. (1948)
12. Mihaescu, M. C. Applied Intelligent Data Analysis: Algorithms for Information Retrieval and Educational Data Mining, pp. 64-111. Zip publishing, Columbus, Ohio (2013)
13. Hall, M. et al. The WEKA data mining software: an update. SIGKDD Explorations, vol 11, 1 (2009)
14. Pentel, A. Age Detector. http://www.tlu.ee/~pentel/age_detector/ (2015)