A Comparison of Different Feature Sets for Age-Based Classification of Short Texts

Avar Pentel

Tallinn University, Institute of Informatics, Estonia

pentel@tlu.ee

Abstract. The aim of current study is to compare two feature sets for age-based classification of short texts as about 100 words per author. Besides widely used n-grams text readability features are proposed as an alternative. By readability features we mean different relative ratios of text elements as characters per word, words per sentence, etc. Support Vector Machines, Logistic Regression, and Bayesian algorithms were used to build models. We found that with 4-year age gap between age groups readability features were more effective for classification than n-grams. With 2-year age gap there was no significant difference in classification results between two feature sets. In both cases combined feature set yield to highest f-scores. Model generated Simple Logistic regression and combined feature set yield to f-score 0.9.

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

1 Introduction

With a widespread of social media, one of the problems is related to anonymity. People can register accounts with false information about themselves. One class of this kind on false information is related to user age. Younger people might pretend to be older in order to get access to sites that otherwise restrict access for them. In the same time older people might pretend to be younger in order to communicate with youngster. As we can imagine, this kind of false information might lead to serious threats, as for instance pedophilia or other criminal activities.

The task of automatic age detection by analyzing written texts belongs to authorship profiling domain. There are two basic types of features that are used for authorship profiling: Content-based features and style-based features. In linguistic terms

those style based features are connected to function words, and content-based features to content words. Tam and Martel's [1] Support Vector Machine model was able to yield a 0.996 f-score when distinguishing teens from adults using word trigram features. One of the main problems of texts authorship profiling, in our case age detection, is that it is almost impossible to classify short texts on the basis of those semantic features. Probability that some sequence of words, even a single word, occur in short text is too low and particular word characterizes better the context [2] than author. Some authors use character n-grams frequencies to profile users, but again, if we speak about texts that are only about 100 words long, these features can also be very context dependent. Some authors [3] argue, that at least 10000 words is needed, other that 5000 [4]. But if we think about business purpose of this kind of age detector, especially when the purpose is to avoid some criminal acts, then there is no time to collect large amount of text written by particular user.

Therefore we propose other set of features that can still characterize author's age in shorter texts, namely text features that are previously used to evaluate texts readability. Texts readability indexes are developed already before computerized text processing, so for example Gunning Fog index [5] takes into account complex (or difficult) words, those containing 3 or more syllables and average number of words per sentence. If sentence is too long and there are many difficult words, the text is considered not easy to read and more education is needed to understand this kind of text. Gunning Fog index is calculated with a formula (1) below:

$$GunningFogIndex = 0.4 \times \left[\left(\frac{words}{sentences} \right) + 100 \times \left(\frac{complexwords}{words} \right) \right] \quad (1)$$

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 actual age of an author. As readability indexes work reliably on texts with about 100 words, these are good candidates for our task. We do not use an actual Gunning Fog Index or any other readability index, but we use the same variables as features in machine learning. As a baseline another n-gram based data set is tested and results are compared.

2 Methodology

2.1 Raw Data

We collected short written texts in average 93 words long. Author of texts are 9-44 years old. All texts in the collections are written in the same language (Estonian). All those texts were digitalized and no errors were corrected.

2.2 Features

In current study we used 3 types of training datasets: with readability features only, with n-grams only and dataset where both readability features and n-grams are present.

Readability features are quantitative data about texts, as for instance an average number of characters in word, syllables in word, words in sentences, commas in sentence and relative frequency of words with 1, 2,…, n syllable. All together 14 different features were extracted from each text plus classification variable (to which age class text author belongs). Complex word in our feature set is loan from Gunning Fog Index [5], where it means words with 3 or more syllables. As Gunning Fog Index is developed for English language, and in Estonian language average number of syllables per word is higher, we raised the number of syllables for complex word to 5. Additionally we count the word complex if it has 13 or more characters.

As n-grams 188 character-bigrams were used. 188 here is not an arbitrary number, it presents all occurred bigrams in whole collection of texts. Character unigrams were not used because there was no significant difference in unigram frequencies between age groups. Character-trigrams were not used because the frequencies were too low.

Both types of features are presented in Table 1.

Table 1. Feature sets. All used readability features are listed here, and out of 188 used character-bigrams top 14 are here for illustration. Underscore (_) symbol substitutes space.

No	Readability features	Character-bigram features
1.	Average number of Character in word	is
2.	Average number of Words in Sentence	as
3.	Complex Words to all Words ratio	_t
4.	Average number of Complex Words in Sentence	li
5.	Average number of Syllables per Word	ta
6.	Average number of Commas per Sentence	in
7.	1 Syllable Words to all Words ratio	te
8.	2 Syllable Words to all Words ratio	ol
9.	3 Syllable Words to all Words ratio	d_
10.	4 Syllable Words to all Words ratio	al
11.	5 Syllable Words to all Words ratio	t_
12.	6 Syllable Words to all Words ratio	_m
13.	7 Syllable Words to all Words ratio	_o
14.	8 or more Syllable Words to all Words ratio.	el

2.3 Data Preprocessing

We stored all the digitalized texts in local machine as separate files for each example. A local program was created to extract all previously listed 14 readability features from each text file, and also 188 character-bigram features. It opened all files one by one; extracted features form each file, and stored these values in a row of a comma-separated file. In the end of every row it stored data about age group. We chose randomly three balanced datasets with 300 records and with different age gaps: 9-15 and 20-44, 9-17 and 20-44, 9-15 and 18-44.

2.4 Machine Learning Algorithms and Technology

For classification we tested six popular machine-learning algorithms:

1. Support Vector Machine
2. Logistic regression
3. Simple Logistic regression
4. Naïve Bayes
5. Naïve Bayes Multinomial
6. Bayesian Logistic Regression

Motivation of choosing those algorithms is based on literature [6,7,8]. The suitability of listed algorithms for given data types and for given binary classification task was also taken in to account. In our task we used Java implementations of listed algorithms that are available in freeware data analysis package Weka [9].

2.5 Validation

For evaluation we used 10 fold cross validation on all models. It means, we partitioned our data to 10 even sized and random parts, and then using one part for validation and other 9 as training dataset. We did so 10 times and then averaged validation results.

3 Results

3.1 Classification of age groups 9-15 and 20-44

With 4-year gap between age groups 9-15 and 20-44 readability features outperformed bigrams with SVM and Logistic regression classifiers (Table 2). Logistic regression and SVM yield both to F-score 0.85. Bayesian classifiers were more effective with bigram features.

Combined dataset with readability features and bigrams yield to little better f-score with standardized data (0.859), but performed poorer with logistic regression. Simple logistic regression was most successful with combined features and yield to f-score

0.9. Bayesian Logistic regression and Naïve Bayes Multinomial fallowed with 0.886 and 0.882 accordingly.

Table 2. Age based classification of short texts written by 9-15 and 20-44 year old authors

Classifier	F-Scores		
	Readability Features	**Character-bigrams**	**Combined**
SVM	0.825	0.803	0.812
SVM standardized	0.850	0.814	0.859
SVM normalized	0.805	0.818	0.845
Logistic Regression	0.850	0.768	0.791
Simple Logistic	0.836	0.818	0.9
Naïve Bayes	0.721	0.791	0.827
Naïve Bayes Multin.	0.799	0.832	0.882
Bayesian Log. Reg.	0.791	0.827	0.886

3.2 Classification of age groups 9-17 and 20-44

With a 2-years age gap between 17 and 20 Most successful classifier-feature set combination we tested was Naïve Bayes Multinomial with bigrams (f-score 0.812). All Bayesian classifiers were less effective with readability features. Logistic Regression classifiers were more successful with readability features. There were no significant differences with SVM, which yield with normalized bigram features to f-score 0.801. Similar result was achieved with standardized readability features and SVM (f-score 0.797).

Table 3. Age based classification of short texts written by 9-17 and 20-44 year old authors

Classifier	F-Scores		
	Readability Features	**Character-bigrams**	**Combined**
SVM	0.793	0.789	0.801
SVM standardized	0.797	0.777	0.793
SVM normalized	0.777	0.801	0.773
Logistic Regression	0.781	0.637	0.746
Simple Logistic	0.777	0.773	0.789
Naïve Bayes	0.703	0.773	0.805
Naïve Bayes Multin.	0.773	0.812	0.848
Bayesian Log. Reg.	0.757	0.809	0.871

Readability features are more effective with SVM if standardized and bigrams if normalized. This leads to the conflict in combined feature set, which cannot improve the results. In both cases – standardization and normalization – combined data set yield to lower results than one of separate data sets. With non-standardized/normalized data and combined feature set classification f-score was 0.801. Using combined dataset,

where the bigram part is normalized and readability features standardized, can do possible improvement.

3.3 Classification of age groups 9-15 and 18-44

Placing age gap two years earlier lead to improvement in results with most classifier-feature combination (Table 4). Similarly to previous data set (Table 3), all Bayesian classifiers performed better with bigrams, while readability features yield to better results with Logistic and Simple Logistic regression.

Table 4. Age based classification of short texts written by 9-15 and 18-44 year old authors

Classifier	F-Scores		
	Readability Features	**Character-bigrams**	**Combined**
SVM	0.803	0.795	0.838
SVM standardized	0.799	0.82	0.863
SVM normalized	0.799	0.829	0.846
Logistic Regression	0.812	0.769	0.786
Simple Logistic	0.812	0.785	0.855
Naïve Bayes	0.706	0.782	0.816
Naïve Bayes Multin.	0.754	0.812	0.85
Bayesian Log. Reg.	0.769	0.825	0.876

With combined feature set most successful classifiers were built by Bayesian Logistic regression, which yield to f-score 0.876, SVM with standardized data (0.863) and Simple Logistic regression (0.855).

4 Discussion

Comparing results of different age gaps and feature sets, we can see (Fig. 1), that readability features yield always to better classification results with Logistic and Simple Logistic regression. Bayesian algorithms perform better with bigram-based features. Support Vector Machines yield formerly [1] to good classification results with n-gram features, but in our study readability features yield to better classifier if age gap was 4 years. With 2-years age gap on 18-19 no significant difference found between best classifiers generated with n-gram or readability features. With other 2-years age gap (16-17) bigram-based features yield to better classifier with Bayesian Logistic Regression, but other Logistic Regression and SVM yield to better results with readabiliy features.

We do not have an explanation why Bayesian algorithms performed always better with bigram-based features. Further study is needed to understand this phenomenon.

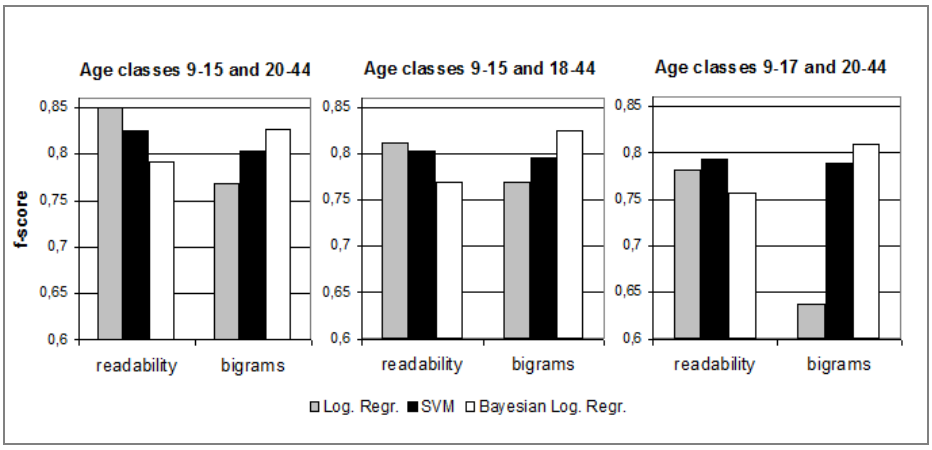

Fig. 1. Effect of features on age based classification with different algorithms and age gaps.

As we can see, readability features depend more on the width of the age gap. Wider age gap yield to better results with readability features, while bigram-based results on first two charts (Fig. 1) are basically identical.

By comparing results of three age groups we can also conclude that position of age gap has influence on classification results. Readability features are more suitable with bigger age gap or with an age gap that is positioned to younger age. With a gap on 16-17 all results were better than with a gap on 18-19. As n-grams depend on vocabulary, and readability features on structure of the texts, we can conclude, that in age 16-17 bigger shift in vocabulary development and in writing will occur, than in age 18-19.

In most cases combined features yield to better results. However this was not the case with Support Vector Machines if age gap was 16-19 or 18-19. Bigram-based features were more effective if normalized and readability features when standardized, therefore it leads to the conflict with combined data set. Standardization of readability part of data set and normalization of bigram part of data set may improve the SVM classifier.

5 Conclusion

Automatic user age detection is a task of growing importance in cyber-safety and criminal investigations. One of the profiling problems here is the amount of text needed to perform reliable prediction. Usually longer texts are needed to make assumptions about author's age. In this paper we tested novel set of features for age-based classification of very short texts (as about 100 words length). Used features are known as text readability features, which are used by different readability formulas, as Gunning Fog, Flesch–Kincaid, etc. These features proved to be suitable for automatic age detection procedure. As a baseline we compared readability features with n-gram-based features, and in many cases readability features yield to better classifica-

tion results. Combined datasets with readability and n-gram features were most successful. Simple Logistic Regression created best model with our data giving 90% classification accuracy.

While this study has generated encouraging results, it has some limitations. As different readability indexes measure how many years of education is needed to understand the text, we can not assume that peoples reading, or in our case writing, skills will continuously improve during the whole life. For most people, the writing skill level developed in high school will not improve further and therefore it is impossible to discriminate between 25 and 30 years olds using only readability features. But these readability features might be still very useful in discriminating between younger age groups, as for instance 7-9, 10-11, 12-13. The other possible utility of similar approach is to use it for predicting education level of an adult author.

Interesting finding in this study was the effect of features on different classification algorithms. Logistic Regressions yield always to better results with readability features, while character bigram-based features were more suitable for Bayesian classifiers. That phenomenon should be explained in future studies.

One limiting factor of current study is the language. For different languages the effect of readability features may be different.

In order to increase the reliability of results, future studies should also include a larger sample. The value of our work is to present suitability of a simple feature set for age based classification of short texts. And we anticipate a more systematic and in-depth study in the near future.

References

1. Tam, J., Martell, C. H. Age Detection in Chat. International Conference on Semantic Computing (2009)
2. Rao, D. et al. 2010. Classifying latent user attributes in twitter, SMUC '10 Proceedings of the 2nd international workshop on Search and mining user-generated contents. pp. 37-44. (2010)
3. 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)
4. 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)
5. Gunning Fog Index. Wikipedia. http://en.wikipedia.org/wiki/Gunning_fog_index
6. Kumar, R., and Verma, R. Classification Algorithms for Data Mining: A Survey. International Journal of Innovations in Engineering and Technology (IJIET) vol. 1 Issue 2 (2012)
7. Wu, X. et al. Top 10 algorithms in data mining. Knowledge and Information Systems. vol 14, 1–37. Springer (2008)
8. Mihaescu, M. C. Applied Intelligent Data Analysis: Algorithms for Information Retrieval and Educational Data Mining, pp. 64-111. Zip publishing, Columbus, Ohio (2013)
9. Weka. Weka 3: Data Mining Software in Java. Machine Learning Group at the University of Waikato. http://www.cs.waikato.ac.nz/ml/weka/