Age and Gender Identification by SMS Text Messages

Ahmad Jamal KHDR
Department of Software Engineering
Technology Faculty, Firat University
Elazig, Turkey
ahmedhurmzi@gmail.com

Cihan VAROL
Department of Computer Science
Sam Houston State University
Huntsville, Texas, United States of America
cvarol@shsu.edu

Abstract — **In this study, age and gender identification are tried to be predicted from SMS text messages. 38,588 preprocessed text messages were tested which were written by native English and Singaporean English students. Naïve Bayes, Support Vector Machine, and J48 Decision tree are applied for gender identification and age range prediction of the author of a given text messages. The test resulted in 70.79% average accuracy for correct age prediction with Support Vector Machine algorithm, and 79.10% average accuracy for correct gender identification via using J48 decision tree.**

Keywordst; Age and gender identification, data preprocessing, J48 decision tree, Naïve Bayes, Support Vector machine, text classification, Weka.

I. Introduction

Age and gender identification depends on understanding a text document's composition through extracting and analyzing the style of writing from the content of the text documents. Lately, analyzing gender and age by electronical and physical text documents has become an important point in our daily life and led up to a rich content of literature of research [1] [2]. Identifying age and gender can be explained in three points of view including attribution of the authorship, verification of the authorship, and characterizing or/and profiling the authorship.

The attribution of authorship includes defining the probable author of a specific text message among a list of familiar writers [3]. Verifying authorship includes checking if a specific writer wrote the text message or not. Profiling or characterizing the authorship includes defining the features such as age and gender of the writer of an unknown text message.

Earlier studies on age or/and gender identification concentrate on general text documents. However, age and gender identification for text messages can play a conclusive role in several criminal cases like terrorist activities and blackmailing. To the best of our knowledge, there are only a few studies conducted on age and gender identification through SMS text messages [4]. Age and gender identification from SMS text messages is complicated and difficult because of the limited characters per page and because the grammatical structure of these text messages may poorly built or written.

II. Earlier Studies

There are various studies have been made for the age and/or gender identification by classification of text documents such as SMS text messages, application review, blog pages, and Facebook posts to name few.

To mention a few, Chaski et al chosen ten writers from their database, the dataset consist of a group of text that contain specific subjects prepared to draw out various records such as love letter, personal letter, and business letter. As for the results they obtained accuracy of author identification at 95.70% [5]. Similarly, Iqbal et al. explained that stylometry-based clustering is useful for profiling writing styles out of unknown e-mail dataset. The experiment procedure applied on real-life e-mail corpus. They achieved 90% accuracy they used k-means to identify authorship of three writes, but later the rate dropped off to 80% after increasing the number of writers to 10 [6].

According to Silessi et al, gender identification includes collecting data, text processing, validation and identifying gender [7]. They used a combination of machine learning algorithms with text processing features for increasing the prediction accuracy of the user gender classification, each data been evaluated by Weka by applying Naïve Bayes, J48, and multi-layer perceptron algorithms, each applied on the data with and without text processing. The reason behind using these algorithms were because statistically they are strong for text classification. After analyzing the results J48 scored better results (accuracy, precision, recall, kappa statistic, and ROC area) compared to Naïve Bayes and perceptron approaches. In addition, Naïve Bayes demonstrated the largest increase in performance after applying text processing.

Iqbal et al suggested Author Miner [8] which comprises an algorithm that gathers repeated syntactical, lexical, content-specific patterns and structures. They used Enron dataset for evaluating their experiments which consisted from 6 to 10 writers, each has 10 to 20 text messages. The accuracy of

978-1-5386-6878-8/18/$31.00 ©2018 IEEE

identifying the authorship changed from 80.5% to 77% after increasing the author's number from 6 to 10.

Argamon et al, studied profiling the author from anonymous text. They used Bayesian Multinomial Regression which in authors' opinion is efficient and accurate. First they started with categorizing the text by taking out all the data and labeling them, each document processed to produce the numerical vector [9]. Later they selected features, and there are two features according to their study; content-based-features and style-based feature. During their study they faced a disadvantage which it is possible that style-based feature markers able to differentiate one class of users from another, and the researchers should worry that content markers may be just artifacts of specific writing situation or experimental setup and may produce positive results that will not be allowed to work out in real-life applications. Therefore, the researchers should be very careful to separate results that has content-based features from those that do not.

Hadjidj et al. used the SVM and C4.5 for authorship identification [10]. In addition, they used three authors from the Enron dataset for evaluation. The classification accuracy for sender identification was 71% and 77%, for sender recipient they achieved 69% and 73%, and for sender-cluster, they obtained 83% in both cases, for SVM and C4.5, respectively.

J. Cheng et al combined the latent semantic indexing method to KNN to predict the gender based on a real life collection of posts on actual blog pages [11]. The researchers did not use only KNN because it is computationally costly due to its lazy learning pattern and it does not perform well when the dimension of feature space is high. According to this paper, KNN is very weak method for classification and it can get better if it is combined with Bag of Words (BoW), Singular Value Decomposition (SVD), Latent Semantic Indexing (LSI), which helps to reduce less informative features. In this study the development and implementation contains five modules; dataset module, data preprocessing module, LSI module, KNN prediction module, and accuracy calculate module. The result for gender identification by using KNN from Facebook posts is 65% and for real life blog posts is %69. However, by applying Naïve Bayes the authors achieved a better result which is %73 for Facebook posts and 71% for real life blog posts.

Simaki et al tried to identify the age of Twitter users out of text [12]. They used a set of text mining, and for classification they estimated recognized and widely used machine learning algorithms such as SVM, J48, RandForest, RandTree, Bayes Net, NBMU, and REPTree so that to show the differences of each one of the algorithms performance. The records clarified that random forest algorithm showed a better performance by yielding 61% of the accuracy.

Montero et al has demonstrated that there are many extensive array of factors that affect writing style such as the gender of the writer and text categories [13], for example word classes, stylometric parameters, and n-grams. In their research, the classification of gender can be noticed as binary classification where particular features which have emotion-based features were chosen from space factorial. For this reason private emotional information that were taken from personal diaries books were used as experiment dataset. As a result, Montero et al yielded 80% accuracy by using Support Vector Machine, but when text documents from blog posts were used, the result changed to 75% while using the same algorithm. This research approved that emotion-based features could be helpful for gender identification.

III. Methodology

For this study, SMS dataset in excel format was downloaded that contained 55,835 short messages in English language. Both native and nonnative speakers in National University of Singapore wrote the messages. Then a systematic procedure has been applied to data to get a significant information about the age and gender of text writers.

Figure I shows the identification process of both age and gender from an SMS text message. The process for both variants followed though collecting data, text processing, and validating to get the outcomes of identification.

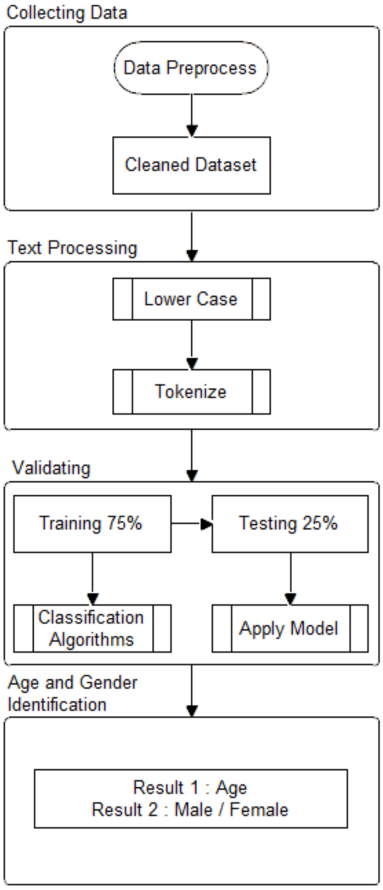

Fig. I Schematic diagram of age and gender Identification Process.

A. *Data preprocessing*

Many unnecessary columns within the dataset was removed because only the column that contains the SMS text message content has been utilized to identify the age and gender. Since, age and gender of one “Sender” is not fixed, “Sender” column was kept to be used for grouping content together.

We removed 10,992records that contain invalid age and gender because it would be useless to be trained on later. This resulted in 41,241 records holding SMS text content, a valid gender, and a valid age.

Nevertheless, since there are some records with valid age and invalid gender they can be used later as training set or testing set for age prediction, on condition to be used for identifying age and not for gender.

Also after removing duplicate messages, 38,600 valid entries left in the corpus. Likewise, twelve messages contained “Vcard” also were removed and thus as final record 38,588 messages were used in our experiments. TABLE I shows the number of instances for both male and females after the preprocessing procedure, while the break down for different age ranges were listed in TABLE .

TABLE I GENDER BREAK DOWN.

Gender	Instances	Percentage ratio
Male	26242	68.01%
Female	12346	31.99%
Total	38588	100%

TABLE II AGE RANGE BREAK DOWN.

Age range	Instance	Instances percentage
16-20	20649	53.51%
21-25	14289	37.03%
26-30	2665	6.91%
31-35	118	0.31%
36-40	595	1.55%
41-45	240	0.62%
46-50	10	0.03%
51-60	20	0.05%

B. *Classification*

There are several classification techniques can be applied to dataset. Some are using 10-fold cross-validation through full dataset to train and testing. The reason behind doing all these tests is to compare between the algorithms and tell which one out of Naive Bayes, Support Vector Machine, and J48 Decision Trees giving the better identification accuracy for age and gender from SMS text messages analyzing. As a result, during all of the tests each one of the SVM, Naive Bayes, and J48 decision tree were run multiple times according to the parameters setting changes.

1) *Age prediction*

In this case, the SVM yielded best accuracy 70.9823% by using the following parameters: IDFT False; TFT False; Output word counts False; Lowercase letters True. However, these parameter settings is not the final result because the average accuracy between all three classifiers is only 61.4894%.

This low record was caused by Naive Bayes’ poor performance. Compare to this record to another set of parameters: which is as the following: IDFT False; TFT False; Output word counts False; Lowercase letters true. The SVM was still the best, at 70.7888%, while the average between classifiers is 65.6681%. Therefore, this parameter setting became the default settings as it recorded the highest score for age prediction by SVM classification algorithm.

2) *Gender Identification*

For gender identification, the SVM yielded best accuracy with 73.5634% by using the following parameters: IDFT False; TFT False; Output word counts False; Lowercase letters True. Nevertheless, not like the age prediction, among all the tests for the gender identification, the following parameter setting yielded best average accuracy, which is as the following: IDFT True; TFT True; Output word counts True; Lowercase letters True, and the average accuracy was 74.6795% but the best algorithm accuracy performance was achieved by J48 at 79.1023, and this setting became the default setting for gender identification.

IV. RESULTS AND DISCUSSION

There are several results in this study while more than one test been made to achieve the satisfying result. The main reason behind each test we did is to increase identification accuracy for each on of age and gender through using only the contents of SMS text messages.

Previously, during the test procedures it was confirmed that 10 fold cross validation” cannot be used instead of splitting testing and training sets willfully. In spite of that, there were six tests successfully run using this method, but to complete each one of these tests we exaggerated in time such as J48 decision tree approach took more than 18 hours. Therefore, 75% of dataset is used for training, while testing set took the rest of the 25%. After repeating the tests according to change the settings of the parameters that used 10 fold cross validation, each one of the scores were matched against those scores that generated from the splitting of 75% of training set and 25% of testing set. Considering that, the scores were not extremely different, and this means:

- Generally, the entire dataset was defined by the 75/25 split.
- it was unnecessary for 10 fold cross validation to be used anymore.

In addition, we have Inverse Document Frequency (IDF) and Term Frequency (TF), which they effect the test score differently because better results can be gained through these parameters, counting on the text nature. Meanwhile, and for the same reasons, the “Convert tokens/words to lowercase” option was diversified.

During some tests, we figured out that:

- For predicting age range we gained better result by using the default settings.
- For identifying gender we gained better result by using the modified settings.
- Without considering of the classifier, most of the tests clarified that diversion to lowercase gained better scores. This may get us to a point that the proper case

existance makes a difference (e.g. YELLING) is quitly enough to misrepresent the score if we didn't put the word on a normal footing.

- The experiments in this study showing that J48 decision tree is the best classifier to identify gender, it has recorded the highest result and Support Vector Machine scored the highest result in age prediction.

  TABLE III showing the final results from our experiments for each used classifiers:

TABLE III FINAL RESULT FOR IDENTIFYING AGE AND GENDER.

Algorithms	Age	Gender
SVM	70.7888%	78.3456%
Naïve Bayes	56.6083%	66.5906%
J48	69.6071%	79.1023%

From the tests, it clarified that Naïve Bayes averaged the fastest to build the model, and J48 average slowest to the build model.

The other tests included comparison between using the (Rainbow) which is a stop word algorithm and both of the (Lovins and Snowball) stemming algorithm. The modified setting of the parameter that recorded the best result in the earlier tests were used for these. Fourteen experiments for gender identification and eleven experiments for age prediction clarified the followings:

- None of the Lovins and Snowball stemming algorithm made any change in the result for any of the SVM, Naive Bayes, and J48 algorithms.
- Using Rainbow (stopword remover), resulted in better performance of Naive Bayes algorithm. Nonetheless, despite of achieving the improvement, but still it was poorer compared to the J48 Decision Trees and the Support Vector Machine.
- It was clear that using the stopword algorithm "Rainbow" dropped the identification accuracy result for the Support Vector Machine, even though it was not that much.
- According to the tests, it seemed that streaming algorithm "Rainbow" has no impact on J48 decision tree algorithm.

After all, we concluded that there is no need for removing stop words, because according to the evidence from each test in regard to interacting with stop words through using the Rainbow algorithm, it suggest that the "StringToWordVector" filter in Weka made satisfactory choices with respect to the words . Obviously, there is no need of removing stop words for this specific text.

In addition, for both Lovins and snowball word stemming filter, it shows the same story. Specifically, applying both Snowball and Lovins stemming filters have very little effect on the age or gender identification.

In addition, that might be caused by the text message's nature the way they been written, because probably that SMS text messages came from a language where the words composed in a way to have less characters.

However, this is normal according to authors of SMS text messages because more often they like to write their messages in less characters usage.

The last group of testing which they are four in total, we used N-Gram modelling for defining if one of the gender or age can be identified more precisely. For each one of the SVM, J48, and Naive Bayes classification algorithms, first the N-Gram modelling was applied, and according to using the previous mention methods they were classified in each one of the tests. To complete the task, by using the entire parameters one will run for each of the classification algorithms, the score were as follows:

- During age prediction, more attributes been created by the N-Gram modelling.
- Curiously, during gender identification, comparing the N-Gram modelling to the other standard Word Tokenizer when they used different parameters, it produced the same number of attributes, where most of them were digrams and trigrams.
- when we used N-Gram modelling, not any one of the algorithms yield better identification accuracy result than before.

During building model, the N-Gram modelling took longer time. Especially the most time consuming process was the final experiment, which performed for predicting age using 1006 attributes and J48; it took more than 18 hours to give the result. This study conducted in English language. In addition it can be enhanced to be used in different languages as well. Also this work can be expanded to improve the success rate when dealing with the structure of shorter messages, e.g. less than 70 characters.

REFERENCES

[1] M. L. Brocardo, I. Traore, S. Saad, and I. Woungang, "Authorship verification for short messages using stylometry," in *Computer, Information and Telecommunication Systems (CITS), 2013 International Conference on*, 2013, pp. 1-6.

[2] A. Narayanan, E. Shi, and B. I. Rubinstein, "Link prediction by de-anonymization: How we won the kaggle social network challenge," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*, 2011, pp. 1825-1834.

[3] S. Nirkhi, R. Dharaskar, and V. Thakare, "Authorship Verification of Online Messages for Forensic Investigation," *Procedia Computer Science,* vol. 78, pp. 640-645, 2016.

[4] C. Sanderson and S. Guenter, "Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 2006, pp. 482-491.

[5] C. E. Chaski, "Who's at the keyboard? Authorship attribution in digital evidence investigations," *International journal of digital evidence,* vol. 4, pp. 1-13, 2005.

[6] F. Iqbal, H. Binsalleeh, B. C. Fung, and M. Debbabi, "Mining writeprints from anonymous e-mails for forensic investigation," *digital investigation,* vol. 7, pp. 56-64, 2010.

[7] S. Silessi, C. Varol, and M. Karabatak, "Identifying Gender from SMS Text Messages," in *Machine Learning and*

Applications (ICMLA), 2016 15th IEEE International Conference on, 2016, pp. 488-491.

[8] F. Iqbal, R. Hadjidj, B. C. Fung, and M. Debbabi, "A novel approach of mining write-prints for authorship attribution in e-mail forensics," *digital investigation,* vol. 5, pp. S42-S51, 2008.

[9] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, "Automatically profiling the author of an anonymous text," *Communications of the ACM,* vol. 52, pp. 119-123, 2009.

[10] R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem, "Towards an integrated e-mail forensic analysis framework," *digital investigation,* vol. 5, pp. 124-137, 2009.

[11] J. Chen, T. Xiao, J. Sheng, and A. Teredesai, "Gender prediction on a real life blog data set using LSI and KNN," in *Computing and Communication Workshop and Conference (CCWC), 2017 IEEE 7th Annual*, 2017, pp. 1-6.

[12] V. Simaki, I. Mporas, and V. Megalooikonomou, "Age identification of twitter users: classification methods and sociolinguistic analysis," in *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing*, 2016.

[13] C. S. Montero, M. Munezero, and T. Kakkonen, "Investigating the role of emotion-based features in author gender classification of text," in *International Conference on Intelligent Text Processing and Computational Linguistics*, 2014, pp. 98-114.