ISSN- 2350-0530(O), ISSN- 2394-3629(P)
Index Copernicus Value (ICV 2018): 86.20
DOI: https://doi.org/10.29121/granthaalayah.v7.i10.2019.408

INTERNATIONAL JOURNAL OF RESEARCH – GRANTHAALAYAH
A knowledge Repository

Science

DEVELOPMENT OF A MACHINE LEARNING ALGORITHM TO PREDICT AUTHOR'S AGE FROM TEXT

Asogwa D.C [1], Anigbogu S.O [2], Anigbogu G.N [3], Efozia F.N [4]
[1, 2] Department of Computer Science, Faculty of Physical Sciences, Nnamdi Azikiwe University Awka, Anambra State, Nigeria.
[3] Nwafor Orizu College of Eduation Nsugbe, Nigeria
[4] Prototype Engineering Development Institute (PEDI), Ilesha, Osun State, Nigeria

Abstract

Author's age prediction is the task of determining the author's age by studying the texts written by them. The prediction of author's age can be enlightening about the different trends, opinions social and political views of an age group. Marketers always use this to encourage a product or a service to an age group following their conveyed interests and opinions. Methodologies in natural language processing have made it possible to predict author's age from text by examining the variation of linguistic characteristics. Also, many machine learning algorithms have been used in author's age prediction. However, in social networks, computational linguists are challenged with numerous issues just as machine learning techniques are performance driven with its own challenges in realistic scenarios. This work developed a model that can predict author's age from text with a machine learning algorithm (Naïve Bayes) using three types of features namely, content based, style based and topic based. The trained model gave a prediction accuracy of 80%.

Keywords: Author Profiling; Machine Learning; Binary Classification; Natural Language Processing.

Cite This Article: Asogwa D.C, Anigbogu S.O, Anigbogu G.N, and Efozia F.N. (2019). "DEVELOPMENT OF A MACHINE LEARNING ALGORITHM TO PREDICT AUTHOR'S AGE FROM TEXT." *International Journal of Research - Granthaalayah*, 7(10), 380-389. https://doi.org/10.29121/granthaalayah.v7.i10.2019.408.

1. Introduction

The problem of identifying the author's age from the text is always of importance as it helps in various fields like forensics and marketing. Author profiling is used to create a profile of an author of a text. Such profile includes the age, gender, native language and the personality traits of the author. In author profiling, linguistic features were used to determine the profile of an author and the most common techniques that are used are different kinds of machine learning techniques (Elias Lundeqvist et al, 2017).

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
Index Copernicus Value (ICV 2018): 86.20
DOI: 10.5281/zenodo.3532022

The author's age is in strong connection with the author's language, as numerous sociolinguistic theories have proven. Depending on the individual's life stage, different linguistic approaches and choices are observed, resulting in the age linguistic variation. A basic principle that differentiates the language of adults from the "teens' language", is that adults use more standard types than adolescents, who prefer non-standard types and generally more unconventional language structures. The author's age from text is a serious sociolinguistics problem that requires a technical attention and this kind of situation requires a model to be built in order to provide a good ground author's age prediction from their text (Dong et al, 2011).

However, computational linguists are challenged with numerous issues in social networks. First of all, little information about the authors' gender, age, social class, race, geographical location, etc., is available to researchers (Herring, 2001). Indeed, most online social networks do not offer open access to the users' profile data. Hence, it is always difficult to collect training or labeled data for this task. Again, communication in online social networks typically occurs via posts on guestbook, blogs, walls, etc. These are usually very short messages, often containing non-standard language usage, which makes this type of text a challenging text genre for natural language processing and machine learning also. Furthermore, given the speed at which chat language has been created generally and continues to develop, especially among adolescents, another challenge in automatically detecting false profiles on social networks is the constant retraining of the machine learning algorithms in order to pick up new variations of chat language usage that are connected to age and/or gender (Herring, 2001).

Therefore, to predict these authors age, text documents can be classified according to a set of predefined classes, using a machine learning technique. The classification is performed based on features extracted from the text documents. These features will be used later to train the classifier. The classifier assigns classes to new data based on the statistics learned from the labeled dataset. Hence, Naive Bayes classifier as a machine learning technique was used.

2. Literature Review

Recently, machine learning approaches have been discovered to estimate the age of an author using text written by the person. This has been modeled as a classification problem, in a similar spirit to sociolinguistic work where age has been investigated in terms of differences in distributions of characteristics between cohorts (Dong Nguyen et al, 2011). In machine learning research, these cohorts have usually been determined for practical reasons relating to distribution of age groups within a corpus, although the boundaries sometimes have also made sense from a life stage perspective. For example, researchers have modeled age as a two-class classification problem with boundaries at age 40 (Garera and Yarowsky, 2009) or 30 (Rao et al., 2010). Another line of work has looked at predicting author's age as a three-class classification problem (Goswami et al., 2009), with age groups of 13-17, 23-27 and 33-42. In addition to machine learning experiments, other researchers have made available statistical analyses of differences in distribution related to age and language and have found similar patterns .For example, Pennebaker and Stone, (2003) analyzed the relationship between language use and aging by collecting data from a large number of previous studies. They used LIWC (linquistic inquiry and word count) (Pennebaker et al., 2001) for analysis. The work showed that with increasing age, people tend to use more positive and fewer

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
Index Copernicus Value (ICV 2018): 86.20
DOI: 10.5281/zenodo.3532022

negative words, more future-tense and less past-tense, and fewer self-references. The results also included, observing a general pattern of increasing cognitive complexity.

Barbieri, (2008) used key word analysis to investigate language and age. Two groups (15–25 and 35–60) were compared. Results showed that younger speakers' speech are characterized by slang and swear words, indicators of speaker stance and emotional involvement, while older people tend to use more modals.

Furthermore, Spiegl et al., (2009) worked on age regression using speech features. Spiegel's system obtained a mean absolute error of approximately 10 years using support vector regression. Van Heerden et al. (2010) explored combining regression estimates to improve age classification. Claudia Peersman et al, 2010 investigated the possibility of automatically predicting age and gender on short chat messages and examined which types of features are most informative for this application. Al-Zuabi et al, (2019) proposed a model that predicts gender and age based on their behavioral services and contact information. They applied different machine learning techniques to provide marketing campaigns with more accurate information about customer demographic attributes. They achieved accuracy of 65.5% for user age prediction.

Morgan et al, (2017) examined the separate and joint predictive validity of linguistic and metadata features in predicting the age of Twitter users. The work created a labeled dataset of Twitter users across three age groups (youth, young adults, adults) by collecting publicly available birthday announcement tweets using the Twitter Search application programming interface and logistic regression.

3. Methodology

For text classification, a supervised machine learning method, Naive Bayes classification was used. Naive Bayes classifier is based on Bayes theorem of calculating posterior probability:

Posterior probability = Prior Probability x Likelihood or for a document d and a class c written as: $P(c|d) = P(d|c)\ P\ (c)/P(d)$

A Naive Bayes classifier makes the assumption that all attributes (input features) are independent of each other.

And for generating set of input features from text, Naive Bayes classification is using text document matrix (bag-of-words model). The model was trained using the content based, style based and topic based features.

4. System Design and Implementation

The rules for author's age prediction from text

The author's age prediction from text of a naive Bayes classifier based on the posterior probabilities can be stated as:

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
Index Copernicus Value (ICV 2018): 86.20
DOI: 10.5281/zenodo.3532022

if $P(\omega$=authortext $|\mathbf{x}) \geq P(\omega$=age$|\mathbf{x})$ classify as authortext, else classify as ham……(1) .
$P(\omega$=authortext $|\mathbf{x}) = P(\mathbf{x}|\omega$=age$)\cdot P(age)P(\omega$=age $|\mathbf{x}) = P(\mathbf{x}|\omega$=age$)\cdot P$(age)……(2)
P^(ω=author)=# of age msg.# of all msg.P^(ω=age)=# of ham msg.# of all msg……(3)

Supposing that the words in every text are conditionally independent based on the naive assumption, two different models can be used to compute the class-conditional probabilities: The models are the Multivariate Bernoulli model and the Multinomial model.

Again, for the evaluation of classification algorithms on the task of author's age prediction, a standard approach was adopted, i.e. data preprocessing, feature extraction and model/classifier training, as illustrated in Figure 4.1 and figure 4.2.

Specifically, each author text from age class post is initially preprocessed. During pre-processing, each post is broken into sentences and each sentence is split into words. Subsequently, the feature extraction approaches are applied in parallel and independently to each post. The text mining, linguistic based and context-based features are extracted; constructing vectors V_{TM}, V_{SL} and V_{CT}, respectively. These features are consequently concatenated to a super vector $V = V_{TM} \| V_{SL} \| VCT$. This results to one feature vector, V, per author's age from text, which is handled by a classification algorithm in order to label each post with an age class

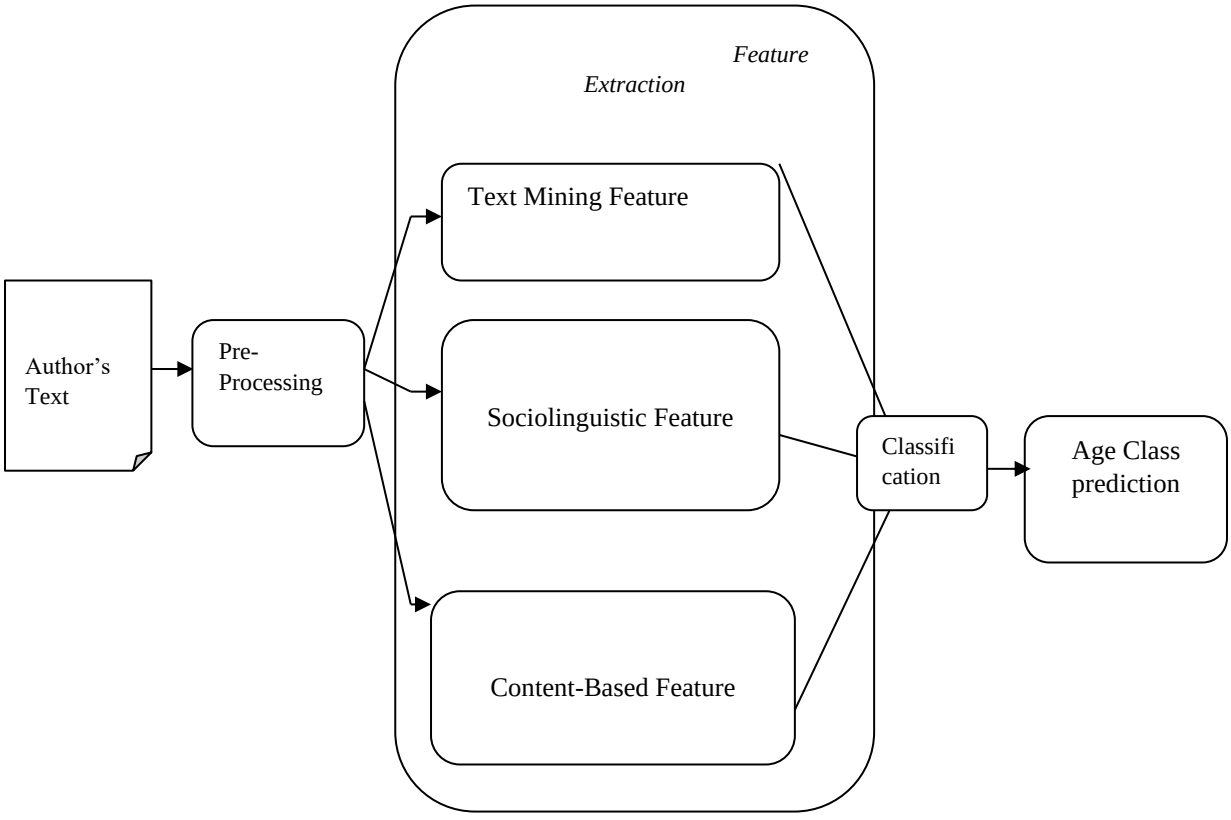


Figure 4.1: Block diagram of the Author's age prediction from text.

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
Index Copernicus Value (ICV 2018): 86.20
DOI: 10.5281/zenodo.3532022

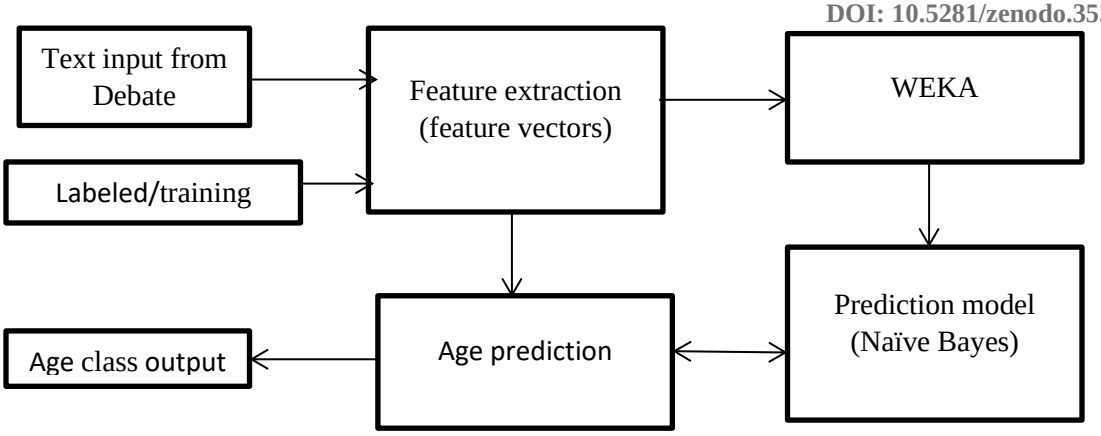


Figure 4.2: System design for author's Age from Text

4.1. Data Set Description

The analysis was based on a particular topic from a given debate. The following topics were given: Declining education standards is caused by laziness on the part of student;

Men always shy away from their marital responsibility;
A good character is better than beauty respectively.

These topics were sent to the student's e mail for responses and classified into three age groups 15–20, 24–30 and 30–34. These data were collated with respect to their age classification and used to train the model for predicting author's age from text.

System Implementation
In the implementation of authors age prediction from text, the training data, test data and 10 cross validation was set up to build the model, that is 90% of training data, testing set of sample data and 10 cross validation. The total sample data used was hundred (100). Weka plugin and java programming language were used. Netbeans IDE was used to integrate the work into the user interface module as shown in figures 5.1 to 5.4.

5. Results and Discussions

Correctly Classified Instances	72	80	%
Incorrectly Classified Instances	18	20	%
Kappa statistic	0.5703		
Mean absolute error	0.2062		
Root mean squared error	0.4481		
Relative absolute error	43.3295 %		
Root relative squared error	91.9106 %		
Total Number of Instances	90		

Correct % = 80.0
Incorrect % = 20.0
AUC % = 0.8742857142857143

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
Index Copernicus Value (ICV 2018): 86.20
DOI: 10.5281/zenodo.3532022

kappa % = 0.570291777188329
MAE % = 0.20623812730234453
RMSE % = 0.4480706881210109
RAE % = 43.32952453733801
RRSE % = 91.91056524141503
Precision % = 0.8135593220338984
Recall % = 0.8727272727272727
fMeasure% = 0.8421052631578948
Error Rate % = 0.2

Table 4.2: The Confusion Matrix

A	**B**	Classified
24	11	**Adult** |= a
7	**48**	**Teenager** |= b

In table 4.2, the confusion matrix is used for evaluation of the result accuracy. The accuracy is calculated by:

Accuracy = (TP + TN) / (TP + TN + FP + FN)
TP = 24, TN = 48, FN = 11, FP = 7
Accuracy = 72/90 = **0.8**
Precision = TP / (TP + FP)
= 24/31
= 0.77419

Table 4.3: Cross validation results

	ADULT	**TEENAGER**	**TOTAL**
ADULT	24	11	35
TEENAGER	7	48	55
TOTAL	31	59	90

From table 4.3, the true positive for adult class is 24 while the true positive for teenager class is 48. The calculations can be seen below.

For Adult Class:
TPR/Recall = TP / (TP + FN) = 24 / (24 + 11) = 24 / 35 = 0.6857
FPR = FP / (TP + FP) = 7 / (7 + 48) = 7 / 65 = 0.1273
PRECISION = TP / (TP + FP) = 24 / (24 + 7) = 24 / 31 = 0.7742
F-MEASURE = 2* PRECISION *RECALL / PRECISION + RECALL
= (2*0.5309) / 1.4599 = 0.7273

For Teenager Class:
TPR/Recall = TP / (TP + FN) = 48 / (48 + 7) = 48 / 55 = 0.8727
FPR = FP / (TP + FP) = 11 / (11 + 24) = 11 / 35 = 0.3143
PRECISION = TP / (TP + FP) = 48 / (48 + 11) = 48 / 59 = 0.8136
F-MEASURE = 2*(PRECISION *RECALL) / (PRECISION + RECALL)

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
Index Copernicus Value (ICV 2018): 86.20
DOI: 10.5281/zenodo.3532022

= (2 * 0.8136 * 0.8727) / (0.8136 + 0.8727)
= 1.4201 / 1.6863 = 0.8421

Figures 5.2 and 5.3 show the receiver operating characteristics (ROC) which evaluates the metrics to check the performance of the model. The Area under the receiver operating characteristics (AUROC) can be shown as 0.91 and 0.87 for the different classes. The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis. An excellent model has AUC near to the 1 which means it has good measure of separability. The TPR is the true positive rate/recall while FPR is the false positive rate/speficity.

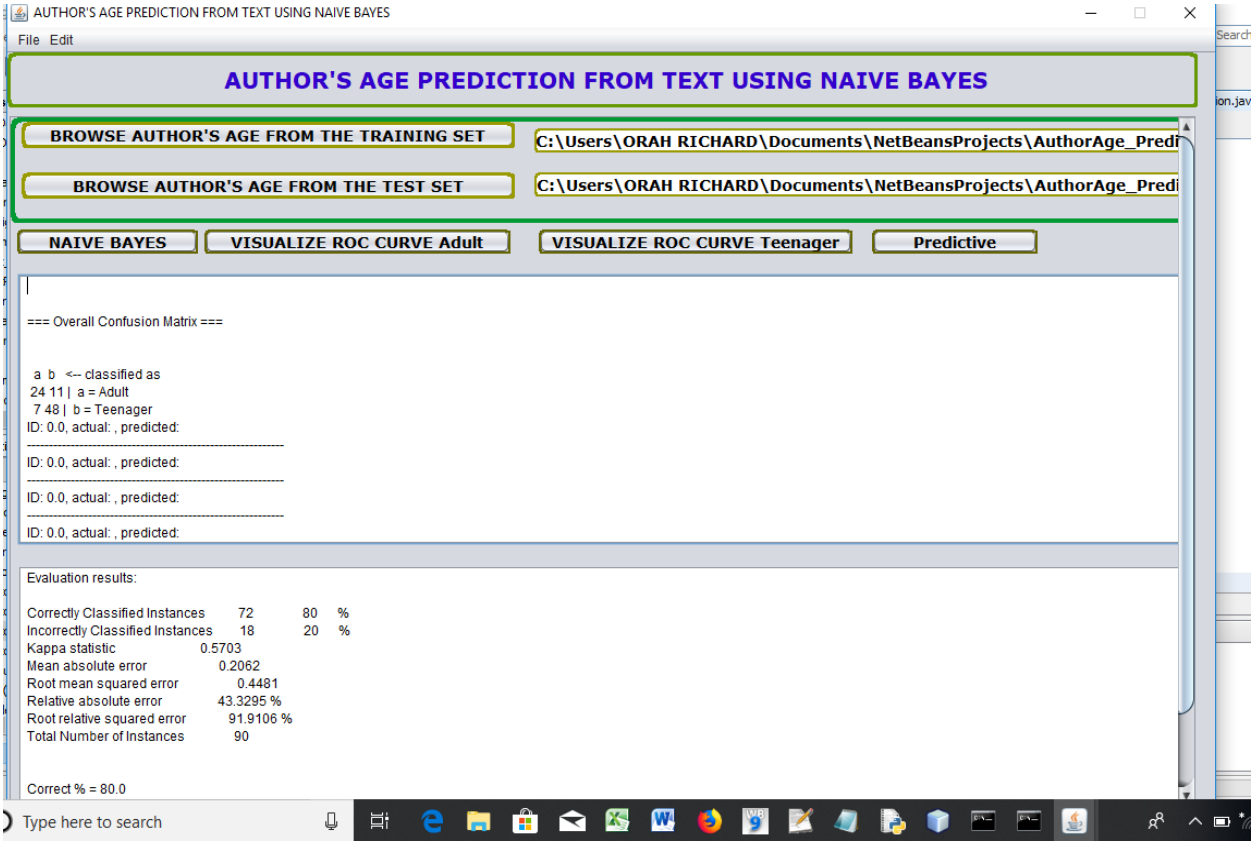


Figure 5.1: Result output

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
Index Copernicus Value (ICV 2018): 86.20
DOI: 10.5281/zenodo.3532022

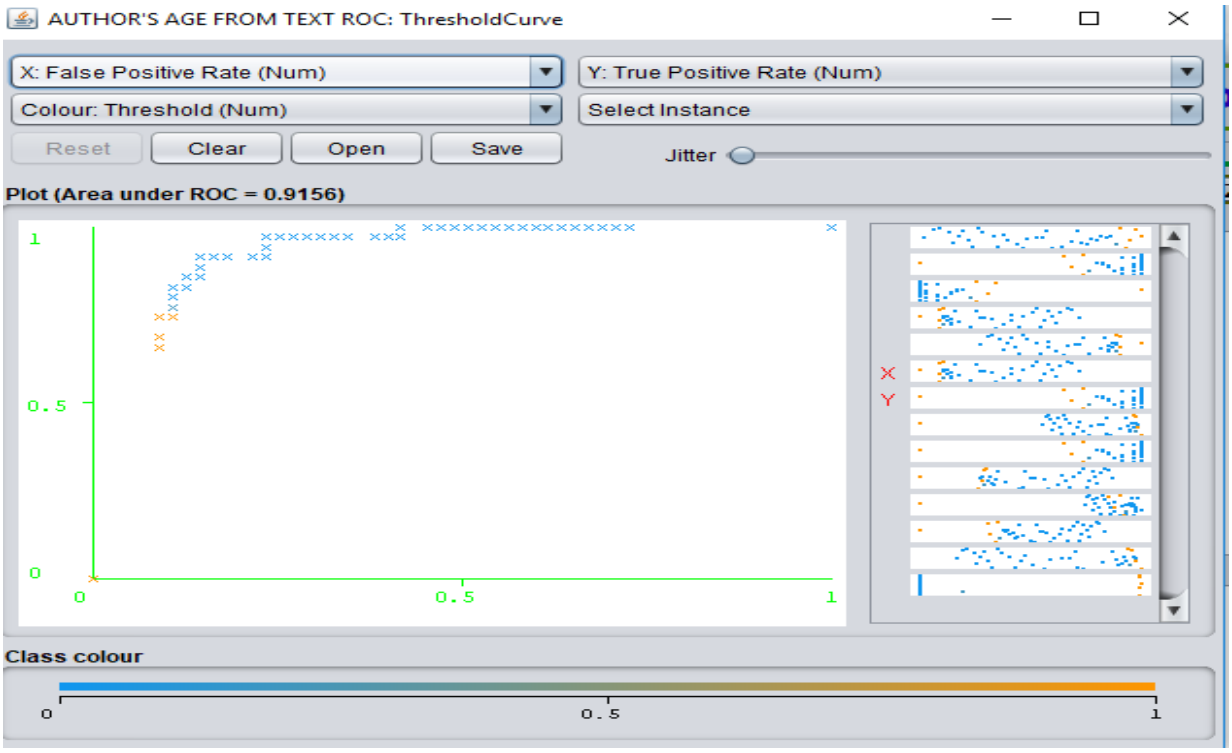


Figure 5.2: The ROC for Adults

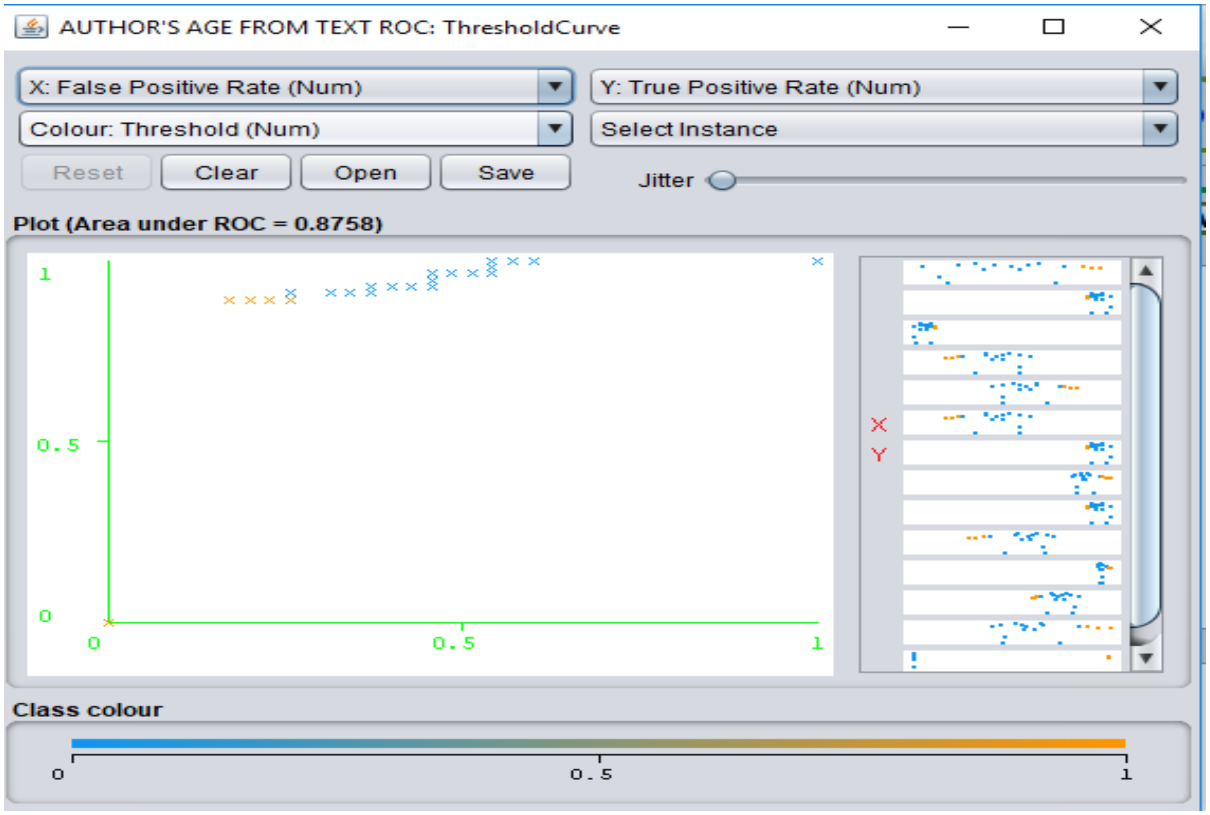
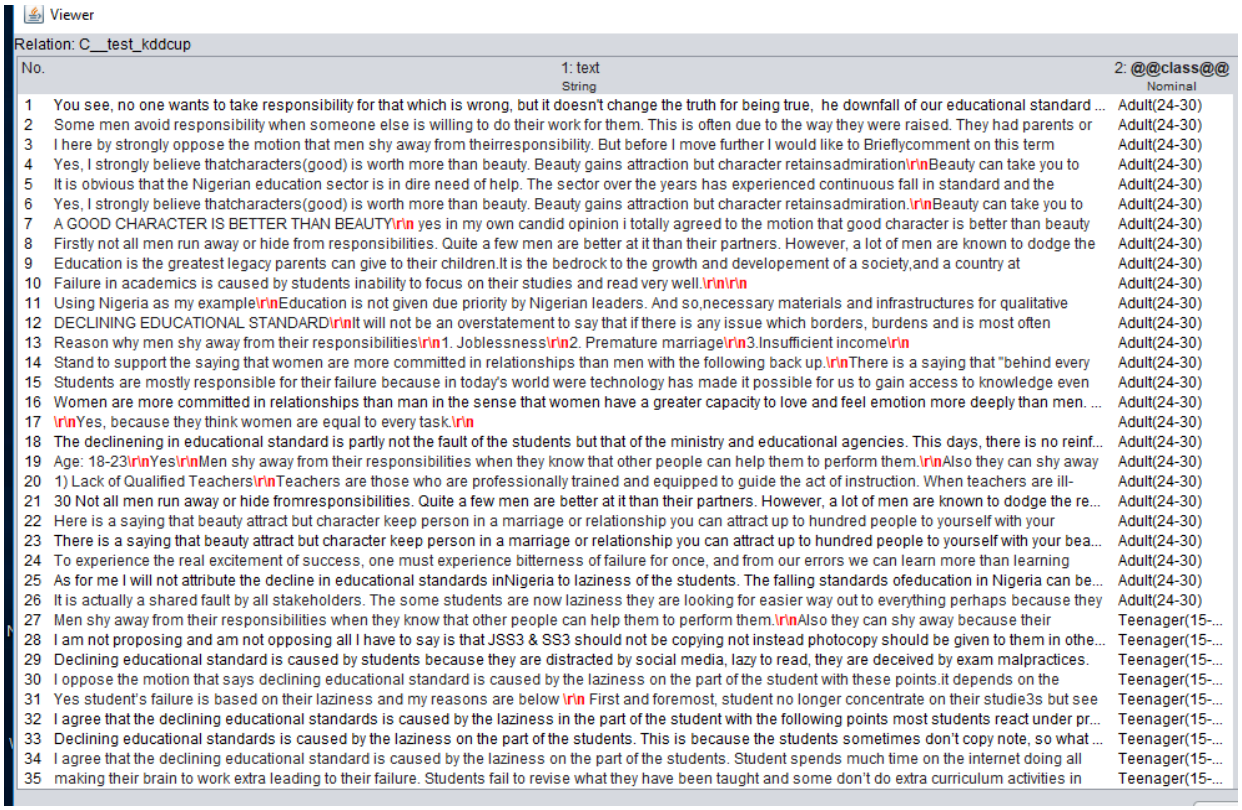


Figure 5.3: ROC for teenagers

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
Index Copernicus Value (ICV 2018): 86.20
DOI: 10.5281/zenodo.3532022



Viewer

Relation: C__test_kddcup

No.	1: text String	2: @@class@@ Nominal
1	You see, no one wants to take responsibility for that which is wrong, but it doesn't change the truth for being true, he downfall of our educational standard ...	Adult(24-30)
2	Some men avoid responsibility when someone else is willing to do their work for them. This is often due to the way they were raised. They had parents or	Adult(24-30)
3	I here by strongly oppose the motion that men shy away from theirresponsibility. But before I move further I would like to Brieflycomment on this term	Adult(24-30)
4	Yes, I strongly believe thatcharacters(good) is worth more than beauty. Beauty gains attraction but character retainsadmiration\r\nBeauty can take you to	Adult(24-30)
5	It is obvious that the Nigerian education sector is in dire need of help. The sector over the years has experienced continuous fall in standard and the	Adult(24-30)
6	Yes, I strongly believe thatcharacters(good) is worth more than beauty. Beauty gains attraction but character retainsadmiration.\r\nBeauty can take you to	Adult(24-30)
7	A GOOD CHARACTER IS BETTER THAN BEAUTY\r\n yes in my own candid opinion i totally agreed to the motion that good character is better than beauty	Adult(24-30)
8	Firstly not all men run away or hide from responsibilities. Quite a few men are better at it than their partners. However, a lot of men are known to dodge the	Adult(24-30)
9	Education is the greatest legacy parents can give to their children.It is the bedrock to the growth and developement of a society,and a country at	Adult(24-30)
10	Failure in academics is caused by students inability to focus on their studies and read very well.\r\n\r\n	Adult(24-30)
11	Using Nigeria as my example\r\nEducation is not given due priority by Nigerian leaders. And so,necessary materials and infrastructures for qualitative	Adult(24-30)
12	DECLINING EDUCATIONAL STANDARD\r\nIt will not be an overstatement to say that if there is any issue which borders, burdens and is most often	Adult(24-30)
13	Reason why men shy away from their responsibilities\r\n1. Joblessness\r\n2. Premature marriage\r\n3.Insufficient income\r\n	Adult(24-30)
14	Stand to support the saying that women are more committed in relationships than men with the following back up.\r\nThere is a saying that "behind every	Adult(24-30)
15	Students are mostly responsible for their failure because in today's world were technology has made it possible for us to gain access to knowledge even	Adult(24-30)
16	Women are more committed in relationships than man in the sense that women have a greater capacity to love and feel emotion more deeply than men. ...	Adult(24-30)
17	\r\nYes, because they think women are equal to every task.\r\n	Adult(24-30)
18	The declinening in educational standard is partly not the fault of the students but that of the ministry and educational agencies. This days, there is no reinf...	Adult(24-30)
19	Age: 18-23\r\nYes\r\nMen shy away from their responsibilities when they know that other people can help them to perform them.\r\nAlso they can shy away	Adult(24-30)
20	1) Lack of Qualified Teachers\r\nTeachers are those who are professionally trained and equipped to guide the act of instruction. When teachers are ill-	Adult(24-30)
21	30 Not all men run away or hide fromresponsibilities. Quite a few men are better at it than their partners. However, a lot of men are known to dodge the re...	Adult(24-30)
22	Here is a saying that beauty attract but character keep person in a marriage or relationship you can attract up to hundred people to yourself with your	Adult(24-30)
23	There is a saying that beauty attract but character keep person in a marriage or relationship you can attract up to hundred people to yourself with your bea...	Adult(24-30)
24	To experience the real excitement of success, one must experience bitterness of failure for once, and from our errors we can learn more than learning	Adult(24-30)
25	As for me I will not attribute the decline in educational standards inNigeria to laziness of the students. The falling standards ofeducation in Nigeria can be...	Adult(24-30)
26	It is actually a shared fault by all stakeholders. The some students are now laziness they are looking for easier way out to everything perhaps because they	Adult(24-30)
27	Men shy away from their responsibilities when they know that other people can help them to perform them.\r\nAlso they can shy away because their	Teenager(15-...
28	I am not proposing and am not opposing all I have to say is that JSS3 & SS3 should not be copying not instead photocopy should be given to them in othe...	Teenager(15-...
29	Declining educational standard is caused by students because they are distracted by social media, lazy to read, they are deceived by exam malpractices.	Teenager(15-...
30	I oppose the motion that says declining educational standard is caused by the laziness on the part of the student with these points.it depends on the	Teenager(15-...
31	Yes student's failure is based on their laziness and my reasons are below \r\n First and foremost, student no longer concentrate on their studie3s but see	Teenager(15-...
32	I agree that the declining educational standards is caused by the laziness in the part of the student with the following points most students react under pr...	Teenager(15-...
33	Declining educational standards is caused by the laziness on the part of the students. This is because the students sometimes don't copy note, so what ...	Teenager(15-...
34	I agree that the declining educational standard is caused by the laziness on the part of the students. Student spends much time on the internet doing all	Teenager(15-...
35	making their brain to work extra leading to their failure. Students fail to revise what they have been taught and some don't do extra curriculum activities in	Teenager(15-...

Figure 5.4: Result output

6. Conclusion and Future Work

A good system for author's age prediction is required in various domains ranging from analyzing sensitive text for national security to commercially important data from various comments and product reviews. This work was able to model the author's age using the writing style and contents of the text. It can be seen that best results were achieved when the context information was used along with the content and style of the texts using a machine learning algorithm, Naïve Bayes for the prediction. Future efforts can be made in this work by introducing inducing sentiment analysis to discover more differences in text written by authors representing different classes. This may yield a much better accuracy rates in identifying the author's profile.

References

[1] Al-Zuabi Ibrahim Mousa, Assef Jafar and Kadan Aljoumaa, 2019, "Predicting customer's gender and age depending on mobile phone data". Journal of big data, https://doi.org/10.1186/s40537-019-0180-9

[2] Charl van Heerden, Etienne Barnard, Marelie Davel, Christiaan van der Walt, Ewald van Dyk, Michael Feld, and Christian Muller. 2010. Combining re-gression and classification methods for improving au-tomatic speaker age recognition. InProc. of ICASSP.

[3] Clauda Peersman, Walter Daelemans & Leona Van Vaerenbergh, 2010 "Predicting Age and Gender in Online Social Networks" Conference'10, Month 1–2, 2010, City, State, Country. Copyright 2010 ACM 1-58113-000-0/00/0010.

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
Index Copernicus Value (ICV 2018): 86.20
DOI: 10.5281/zenodo.3532022

[4] Rao Delip, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010."Classifying Latent User Attributes in Twitter". In: Proceedings of the 2Nd International Workshop on Search and Min-ing User-generated Contents. SMUC '10. Toronto, ON, Canada: ACM,2010, pp. 37– 44. url: http://doi.acm.org/10.1145/1871985.1871993 (cit. on p. 4).

[5] Dong Nyuyen, Noah A., Smith Carolyn P., & Rose, 2011, Author age prediction from text using linear regression, Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 115–123,Portland, OR, USA, 24 June 2011.c©2011 Association for Computational

[6] Elias Lundeqvist & Maria Svensson, 2017, "Author profiling: A machine learning approach towards detecting gender, age and native language of users in social media" M Sc thesis, Department of information technology, Uppsala, http://www.teknat.uu.se/student, UPTEC IT 17013

[7] Federica Barbieri. 2008. Patterns of age-based linguistic variation in American English. Journal of Sociolin-guistics, 12(1):58–88.

[8] Goswami, Sudeshna Sarkar, and Mayur Rustagi.2009. Stylometric analysis of bloggers' age and gen-der. InProc. of ICWSM.

[9] Herring, S. C. 2001. Computer-mediated discourse. In Schiffrin, D., Tannen, D., and Hamilton, H.E. (eds.), The Handbook of Discourse Analysis. Blackwell, Malden, Massachusetts, USA, 612 -634. DOI=10.1111/b.9780631205968.2003.x

[10] Pennebaker James W and Lori D. Stone. 2003. Wordsof wisdom: Language use over the lifespan. Journalof Personality and Social Psychology, 85:291–301.

[11] Pennebaker James W., Roger J. Booth, and Martha E. Francis, 2001.Linguistic Inquiry and Word Count (LIWC): A Computerized Text Analysis Program.

[12] Nikesh Garera and David Yarowsky. 2009. Modeling la-tent biographic attributes in conversational genres. InProc. of ACL-IJCNLP. Sumit

[13] Morgan-Lopez AA, Kim AE, Chew RF, Ruddle P (2017) Predicting age groups of Twitter users based on language and metadata features. PLoS ONE 12(8): e0183537. https://doi.org/10.1371/journal.pone.0183537

[14] Werner Spiegl, Georg Stemmer, Eva Lasarcyk, Varada Kolhatkar, Andrew Cassidy, Blaise Potard, StephenShum, Young Chol Song, Puyang Xu, Peter Beyer-lein, James Harnsberger, and Elmar N¨oth. 2009. Ana-lyzing features for automatic age estimation on cross-sectional data. InProc. of INTERSPEECH.

*Corresponding author.
E-mail address: dc.asogwa@ unizik.edu.ng/so.anigbogu@unizik.edu.ng/anigbogugloria@ yahoo.com/ efozia23@ yahoo.ca