Stylometry Based Authorship Identification on ChatBot Corpus

Vishal Paranjape[#1], Dr. Ravindra Patel[#2],Dr. Nishchol Mishra[#3]

Ph.d Scholar[#1], Dept. of Computer Application[#2], SOIT [#3]

RGPV, Bhopal, India

Abstract - In today's scenario authorship analysis technique has a great relevance as it is visualized in three perspectives namely Authorship Profiling, Authorship Identification & Plagiarism Detection. The objective of this paper is to provide a review on the various studies conducted on Authorship styles. The present paper also incorporates the use of Chatbots for developing conversation based log file and to propose the technique for identifying the actual author using the log file of ChatBot. There are wide applications of authorship profiling where it is mostly used in marketing, security and forensics. Predicting authors age, gender and personality traits on the basis of writing style of authors play a prominent role in the field of forensic science. During the past some areas there is tremendous development in this field with the help of machine learning, natural language processing and information retrieval. This paper is based on survey to predict the approaches for authorship attribution for both text representation and text classification by examining their characteristics and the most challenging field of artificial intelligence which makes use of ChatBots for communication.

Keywords: Chatbot, machine learning, personality trait, author profiling, gender prediction, age prediction.

1. INTRODUCTION

The most straight away technique adopted in author attribution comprise of finding the actual author of a given document and we are asked to determine which of the small set of candidates the actual author of is given text. The root of authorship analysis lies in a linguistic research area called stylometry, which refers to statistical analysis of literally style [1]. Authorship analysis studies can be classified in three ways [2]. They are:

***Authorship attribution or identification*:** it determines if a particular text being written by an author.

***Authorship profiling or characterization*:** It diagnoses the profile or characteristic of an author that produced a given piece of work.

***Similarity detection*:** It compares multiple pieces of work & determines its genuineness. It is mostly used for plagiarism detection.

Crime investigation to identify the culprit is done by the help of Author profiling on the basis of their writing style. Today a lot of crime is increasing due to enhancement of social networking which is helpful in increasing crimes like public harassment, fake profiles, defamation, blackmailing etc. By knowing the writing style of author it is easy to catch the culprit of a given offense. In the field of marketing author profiling helps in identifying the genuineness of the review or feedback given by the consumer on a particular product that helps in making new and better business decisions according to the needs of the consumer. The present scenario shows that things have changed a lot with respect to authorship attribution. Information retrieval plays a prominent role in this era due to vast amount of electronic text available through internet media and this is the reason for development of natural language processing. Due to these advancement there is development of techniques of authorship attribution technologies as described below:

- For representing and classifying large volumes of text information retrieval research was developed.
- Powerful machine learning algorithms is available for handling sparse and multidimensional data.
- In order to represent the style and analyze text efficiently some tools were developed by NLP.

1.1 Authorship Analysis

This technique is concerned with finding the real author of an anonymous document .This technique comprise of the techniques of feature extraction and data cleaning followed by normalization and feature extraction. Stylometric features are used for calculating feature values [3]. The feature which is extracted is classified into training and testing sets. Testing set is used to validate the developed model and Training set is used to develop a model.

1.2 Authorship Characterization

Sociolinguistic attributes like age, occupation, gender and educational level of potential author of an anonymous document is detected using this characterization. [4][5]

2. LEXICAL FEATURES

These consist of the word unigrams, bigrams and trigrams, which are commonly used in an author's profile.

Twitter Style: Most of the style of authors is recognized with the help of the following features used in tweets like number of words, characters, question marks, exclamation marks, hash tags, average length of tags.

Familial Tokens : Some tokens are helpful in diagnosing whether the author of a matter is a male or a female , as females mostly use the words my hubby, my husband, my boyfriend etc. whereas males mostly use the matter my girlfriend, my wife etc. These words in tweets are quite reasonable to predict the gender of author.

LDA Topics: In this era the concept of LDA topics are widely used to predict age, gender, personality of authors researchers.

2.1 Content-specific features

Few keyword and terms are used by it for characterizing certain content-specific features and discussion forums or interest groups by a few keywords or terms [6]. Authors in manually analyze and observe historical messages and identify 11 key words as content-specific features particularly for English "for-sale" online messages (sale, windows, Microsoft …).

On the basis of above discussion soft most useful stylometric features are given as below in the table 1.

Table 1 : Stylometric features

Lexical features (F1)	
Character based features	Characters count (c)
	Total number of alphabetic characters/C
	Total number of upper-case characters/C
	Total number of digit characters/C
	Total number of white-space characters/C
	Total number of tab spaces/C
	Frequency of letters (26 features) A-Z
	Frequency of special characters ~, @,#,$,%,^,&,*,_ etc.
Word – based features	Total number of words (M)
	Total number of short words (less than 4 characters)
	Average word length
	Average sentence length in terms of character
	Yule's K measure
Syntactic Features (F2)	Frequency of punctuation ",", "?", "!" , ":", ";"
	Total number of lines
	Total number of sentences
Structural Features (F3)	Total number of paragraph
	Number of sentence per paragraph
	Number of words per paragraph
	Use e-mail as signature
	Use telephone as signature
	Use URL as signature

3. RELATED WORK

Nawaf Ali et.al. [7] investigated the chatbot linguistic style and its effect on authorship attribution. The study shows that chatbots did show a behavioral drift in their style. Source of data was Loebner prize from where data was downloaded and in order to clean the files Perl script was used and split each chat into two text files one for chatbot under study and other for human judge.

```
Copyright (C) 2003 Cambridge Center for Behavioral Studies. All rights reserved.
ALICE Silver Edition conversing with Judge 06

2003-10-18-15-40-46 PROGRAM: Hello?  Is anyone there?
2003-10-18-15-41-00 JUDGE: Hi. Who are you?
2003-10-18-15-41-05 PROGRAM: Hello there. I am Alice. Who are you?
2003-10-18-15-41-22 JUDGE: I'm Marilyn. What are you doing here?
2003-10-18-15-41-27 PROGRAM: Hi there Marilyn. I am chatting with people online.
What are you doing?
2003-10-18-15-41-48 JUDGE: Well at the moment I'm talking to you Alice?
2003-10-18-15-41-53 PROGRAM: Well that's okay. Sounds like fun.
2003-10-18-15-42-11 JUDGE: It could be. What shall we chat about then?
2003-10-18-15-42-16 PROGRAM: Under what circumstances? Quite a lot of things.
2003-10-18-15-42-51 JUDGE: As we have only just met perhaps we should just get to
know each other
2003-10-18-15-42-56 PROGRAM: Do you mean your name is we have only just met perhaps
 we should just get to know each other?
2003-10-18-15-43-15 JUDGE: No. My name is Marilyn
2003-10-18-15-43-20 PROGRAM: What is your real name? I am always glad to make new
friends, Marilyn.
2003-10-18-15-43-43 JUDGE: That is my real name. Are you really called Alice?
2003-10-18-15-43-48 PROGRAM: Who gave you your real name? I don't know whether or
not I am called. I am a Mammal.. Really.
2003-10-18-15-44-19 JUDGE: My parents gave me my name.
2003-10-18-15-44-24 PROGRAM: Such an original name!
2003-10-18-15-44-42 JUDGE: Thank you. Who are you called after?
2003-10-18-15-44-47 PROGRAM: you are quite welcome. Human beings. I have never been
 asked that before. How did you hear about Alice?
```

Fig 1: Conversation of Program and Judge

RapidMiner platform was used for conducting our experiment and a model was built for authorship identification using support vector machine that will accept the training text and create a word list and a model and then this word list and model will be implemented on the test text which is the data from Loebner prize

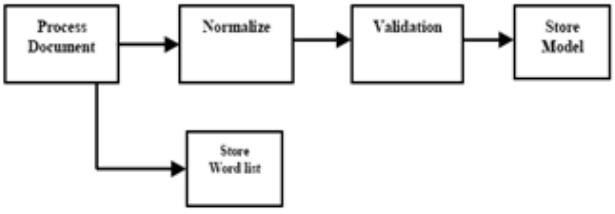

Fig 2: Training model using RapidMiner

Now we used the saved wordlist and model as input for the testing stage.

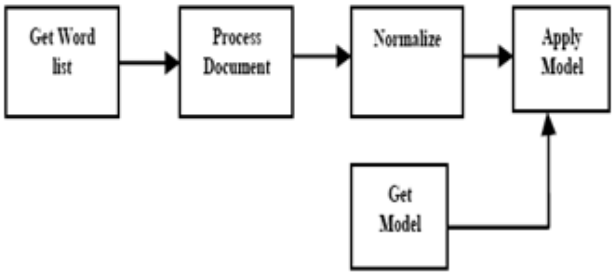

Fig 3: Testing stage using RapidMiner

By using these experiments we can reveal that some chatbots do change their style depending on intelligent algorithms used in initializing conversations.

V. Roman et.al.[8] told it as an area of investigation as intelligent authors have not been profiled based on their linguistic behavior. Collected data comes from chatbot logs between chatbots and human users. An application was developed to connect chatbots and collect data from logs. After cleaning the data being collected there were two applications used the first was Java Graphical Authorship Attribution Project (JGAAP) and other was stylometry. JGAAP has given ability to test each feature performance on our data.

T. Raghunadha Reddy et.[9] Al. suggested technique to identify the writing style characteristics of authors. This survey broadly focuses on predicting the demographic of authors such as personality traits, age, gender etc. based on the text corpus written by various authors.

Based on the extensive literature survey we can adopt the technique for identification of author using a chatbot Verbot 5 which is an open source application. We can make multiple dialog conversation between user and Verbot (ChatBot). The dialog can be extracted using log file available.

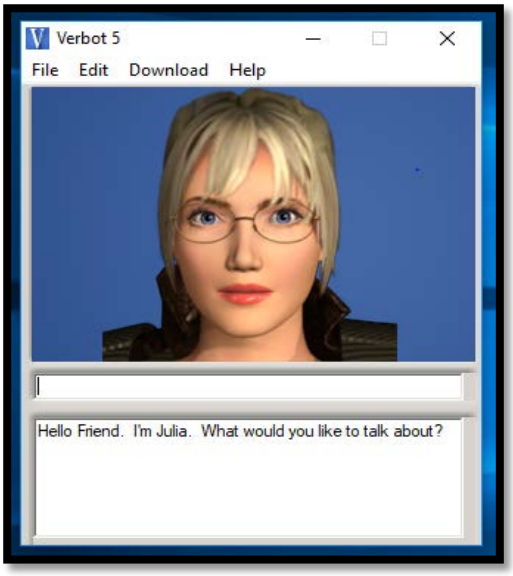

Fig 4: ChatBotWindow

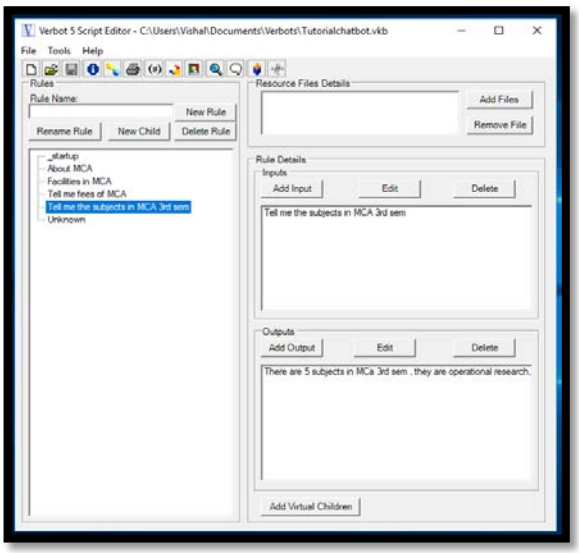

Fig 5: ChatBot Script Editor

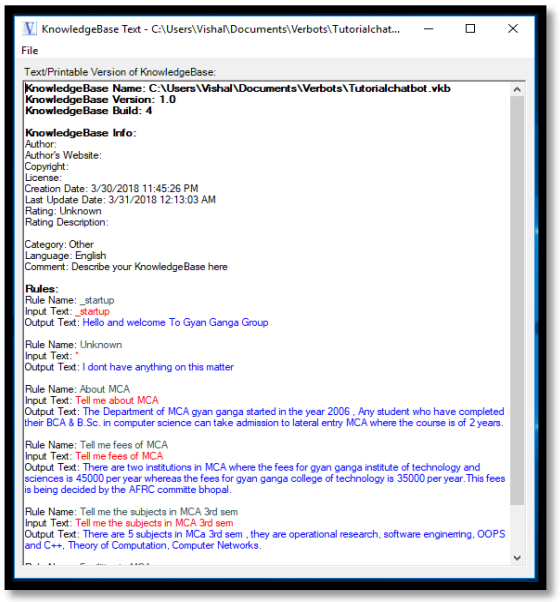

Fig 6: KnowledgeBase

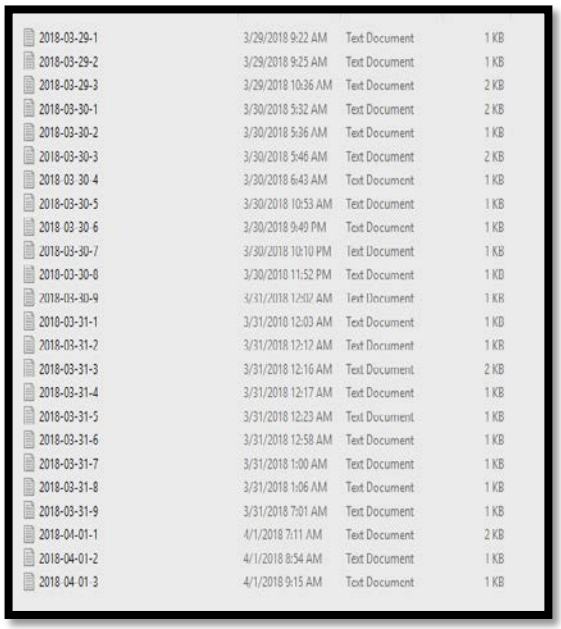

Fig 7:ChatBot Corpus Log File

We can apply the model which will accept the training text and create a word and create a model using Support Vector Machine (SVM) & a model. Our experiments model will output the confidence reflecting how confident we are that the chatbot is identified correctly. The chatbot with highest confidence value is the predicted bot according to the model.

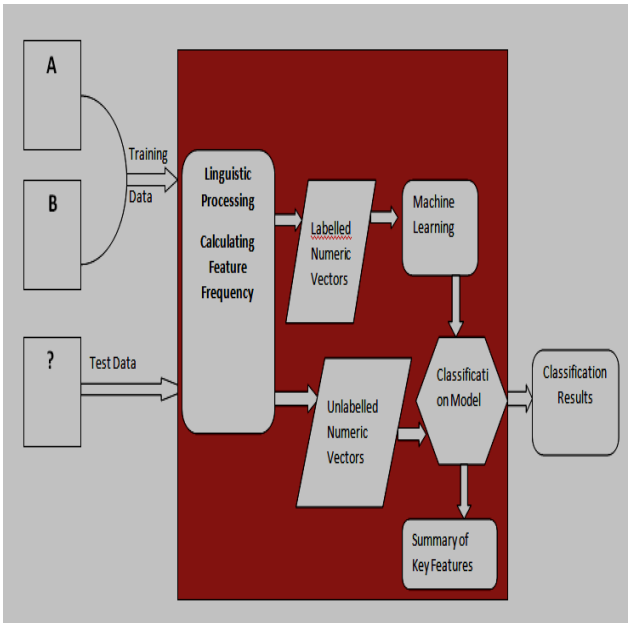

Fig 8: Authorship profiling using Machine Learning

The above diagram depicts the architecture for authorship profiling using machine learning. The labeled document such as gender are used as training data; which are tagged and processed linguistically and calculated, giving a numeric vector for each individual text, labeled with the text's correct authorship label. Classification model is created by machine learning method which is then applied to vectors computed from unlabeled test documents – classification accuracy gives a measure of how effective the technique is, while the most significant features for classification give a rough characterization of the linguistic difference between given author types.

4. CONCLUSION

We have shown how the right combination of linguistic features and machine learning methods enables an automated system to effectively determine several such aspects of an anonymous author; it is likely that other important profile components (such as educational background or other personality components) can also be extracted using such techniques, given appropriate training material. An important open research question, however, is the extent to which variation in genre and language might affect the nature of the models that can be used to solve various aspects of the profiling problem.

REFERENCES

[1] H. Chen, Z. Huang, J. Li, R. Zheng , "A framework for authorship Identification of Online Messages: writing-Style features and classification Techniques", JASIST, 2006.

[2] H. Chen, Z. Huang, Y. Qin, R. Zheng, "Authorship Analysis in Cybercrime Investigation" (Eds.): ISI 2003, LNCS 2665, pp : 59-73, 2003.

[3] F. Iqbal, "Messaging Forensic Framework for Cybercrime Investigation", A Thesis in the Department of Computer Science and Software Engineering - Concordia University Montréal, Canada,2011.

[4] D. Abbott, M.J. Berryman, S. Jain, T.J. Putnins, D.J. Signoriell, "Advanced text authorship detection methods and their application to biblical texts. The International Society for Optical Engineering", pp : 1-13,2006.

[5] F. Iqbal, "Mining writeprints from anonymous e-mails for forensic investigation". DigitInvestig,doi:10.1016/j.diin.2010.03.003,2010.

[6] H. Chen, Z. Huang, J. Li, R. Zheng "A framework for authorship Identification of Online Messages: writing-Style features and classification Techniques", JASIST, pp : 378-393,2006.

[7] N. Ali, M. Hindi, Roman V. Yampolskiy," Evaluation of Authorship Attribution Software on a Chat Bot Corpus", 978-1-4577-0746-9/11/$26.00 ©2011 IEEE.

[8] D.Schaeffer, Roman V. Yampolskiy," Linguistic Profiling and Behavioral Drift in Chat Bots".

[9] T. Raghunadha Reddy, B. Vishnu Vardhan, P. Vijayapal Reddy," A Survey on Authorship ProfilingTechniques",http://www.ripublication.com, 2016.