A Framework for Authorship Identification of Questioned Documents: Forensic and Linguistic Convergence by Sresha Yadav and Smita Jha

A Framework for Authorship Identification of Questioned Documents: Forensic and Linguistic Convergence

Sresha Yadav * and Smita Jha **

*Research Scholar, Department of Humanities & Social Sciences Studies, Indian Institute of Technology, Roorkee-247667, India, E-mail: sresha85@gmail.com

**Assistant Professor, Department of Humanities & Social Sciences Studies, Indian Institute of Technology, Roorkee-247667, India, E-mail: smitaiitr@gmail.com

Abstract

Any questioned document examination in India is limited in its scope of authorship identification when the investigators are unable to obtain the criminal's standard handwriting specimen. Every person's style of handwriting is unique and has its own personalized touch. It is because of this reason that handwriting is very difficult to disguise and forge, making handwriting analysis an effective tool for incriminating a suspect. The application of linguistic tools such as stylistics can support Forensic techniques in eliminating this limitation. Using three types of writing style features such as lexical, syntactic, and structural features, we have attempted to develop a framework for authorship identification by analyzing these features in the questioned text as well as comparing its results with those of known authors based on forensic handwriting principles. This study will help in not only authorship identification but will exclusively help in linking the criminal's motive (*mens rea*- guilty mind) with the evidence and crime. Therefore, this study uses a convergence framework using forensic and linguistic techniques for authorship identification.

Keywords: Questioned document, Linguistics, Forensic handwriting principles

A Framework for Authorship Identification of Questioned Documents: Forensic and Linguistic Convergence by Sresha Yadav and Smita Jha

Introduction

In the present scenario with the emergence of new and sophisticated technologies, crime in India has taken a new ramped. In the field of Questioned Document examination, investigating personnel's sometimes faced problems due to the non-availability of standard specimens from the suspects of crime. Thus, one of the major challenges in front of the law enforcement agencies is to develop an investigative tool which in addition to the traditional approach of authorship identification also increases the effectiveness of the identification procedure. Application of linguistic techniques for authorship identification will augment the efficacy of examination procedure even if the availability of standard samples is limited or nil. To fix the authorship of any Questioned Document is important in the context of 'white collar crimes" i.e. crime involving forgery, identity thefts, counterfeiting, anonymous letters such as threatening or suicide note. The examination of these types of documents is vital because they are produced directly in the court of law as a connecting link between the perpetrator and the committed crime. For the analysis of a these type of documents , it is important to study the written language of the perpetrator, as the vocabulary usage in the text may yield "signature" words unique to the offender(Douglas et al, 1986).

Handwriting Analysis and Authorship Identification

In India, traditional approach of handwriting analysis is been employed for authorship identification. The process involves extracting certain characteristic features present in a text to determine the authorship (Zheng et al, 2006). Class and individualistic characteristics are used for this purpose of examination. On the basis of handwriting principles, these characteristics are determined when compared to standard specimens obtained from the suspect. The main principles of handwriting analysis is based on presumed facts that the writing of every individual is personal to self only , two different persons can't write in similar manner, presence of natural variations in writings of a same

A Framework for Authorship Identification of Questioned Documents: Forensic and Linguistic Convergence by Sresha Yadav and Smita Jha

person, disguised writing always lead to inferior quality of writing and writing of every individual is a response to brain stimuli so it is a kind of brain writing which is an acquired skill and helps to determine the individuality of the person through writing features. The writing practices we learn during our time at school are very difficult to lose, as we get used to the particular way that we hold a pen, shape the letters we write and how we space our words and lines. These are some of the factors that prove useful during the analysis of a document. Investigators analyze these aspects of suspicious documents i.e the printing style, paper and ink, all of which help to identify a forged letter. The handwriting section of forensic science involves the comparing and authentication of written documents such as *ransom* notes, forged contracts, forged wills, fake ID's and passports and any other form of writing or printed material. The analysis of someone's handwriting is most commonly used to prove that two documents were written by the same person. When looking at a person's handwriting, the examiners usually look for personalized characteristics under four areas including line quality, form, content and arrangement. The form of writing involves examining the shape of singular letters and identifying if the slant is in a certain direction, the size and how they are connected with the next letter. Unusual characteristics, such as the use of a plus sign or the ampersand (&) are also noted. Examining the content of written and printed papers is done to identify similarities between punctuation, spelling, grammar, vocabulary and paragraph phrasing.

Document examiners compare unidentified documents with a 'standard', a sample from a suspect. A standard is usually produced by the suspect under supervision. Even under supervision, the suspect still has the chance to disguise their handwriting, which is why investigators then have to collect other standards of casual handwriting from a suspect. The casual handwriting is undisguised and can therefore be compared with the unknown sample either with words that match or letter-by-letter.

A Framework for Authorship Identification of Questioned Documents: Forensic and Linguistic Convergence by Sresha Yadav and Smita Jha

Scope of Linguistics in Authorship Identification

The language used in the questioned text is of great importance because each and every individual possesses certain characteristic features which will help to identify the geographical origin, age, occupation, sex, education, and religious background by the study of the language used in the text. The study by Koppel, Argamon, and Shimoni (2002) provided evidence for the fact that how the language used by male and female study group varies with respect to the use of pronouns. Corney, Vel, Anderson, and Mohay (2002) points out that how the educational background of any person is reflected in the language which is used by the person. In recent years many researchers explored the area of authorship identification in electronic messages and proposed different classification techniques with multidisciplinary approach to identify the author of the unknown text with greater accuracy. The main problem in any questioned text is the factor of anonymity, the individual tried to refrain the basic identity information i.e. gender, age, occupation etc (Zheng et al, 2006). The importance of the linguistic analysis in the examination of the questioned text provides information with respect to suspect age, gender, race, occupation, and educational background even if the availability of the standard writing specimen is limited or nil. Application of linguistics tools also extends the principles of psycholinguistics techniques to sketch the offender profile which can be used to identify anonymous letters writers (Casey Owens, 1984) and any person who make written or spoken threats of violence (Miron and Douglas, 1979). The field of psycholinguistics is concerned with the relationship between linguistics and the psychological processes underlying the. Linguistic features also help in linking the individual motive (*mens rea*- guilty mind) by analyzing character styles and personality traits to understand or to predict criminal behavior
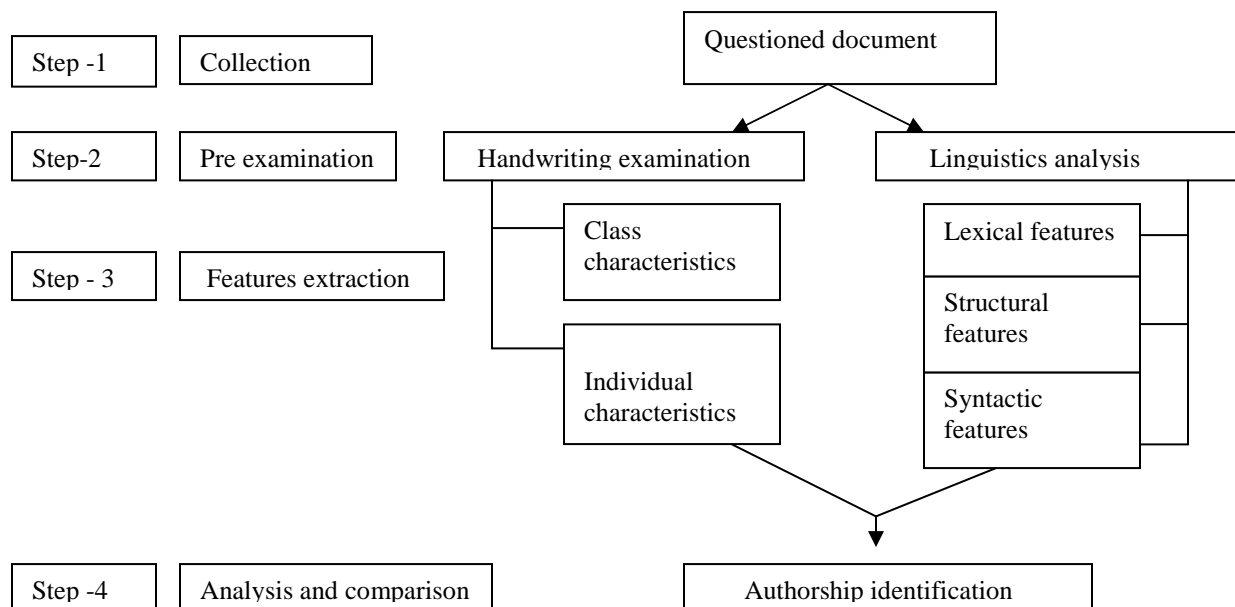
Methodology

The main purpose of this study is to develop a framework utilizing the techniques of linguistics and forensics which not only helps in authorship identification but also to

A Framework for Authorship Identification of Questioned Documents: Forensic and Linguistic Convergence by Sresha Yadav and Smita Jha

establish the criminal motive.   On the basis of handwriting analysis individual characteristics i.e. size of the letters, arrangement, spacing between the letters, words and lines, pen pause, pen lift, hesitation and connecting strokes; class characteristics i. e. pictorial effect, style, movement of writing, writing speed, alignment of letters, and line quality of the written text is to be determined and compared with the results obtained from linguistic analysis. Based on the review of earlier related studies and analysis of the questioned text of written document or computer generated document, we considered three types of linguistic feature set: lexical, syntactic and structural features.

In this study we included lexical feature used in de Vel (2000), vocabulary richness(Yule, 1938); syntactic features, including function words (Mosteller and Wallace, 1964), part of speech incorporation (Stamatatos et al,2001), punctuation characteristics (Baayen et al., 2002) . The structural features represent the writes style of organizing the layout of writing text (Zheng, 2006).

Figure I. Framework proposed for Authorship Identification

A Framework for Authorship Identification of Questioned Documents: Forensic and Linguistic Convergence by Sresha Yadav and Smita Jha

Conclusion and Future Challenges

In this present study, we proposed a framework for authorship identification of questioned text by adopting the linguistic as well as forensic approaches and techniques. We believed that the proposed framework has the potential to assist in authorship identification and also to sketch the linguistic behaviour profile of the suspect. As this study is in nascent phase its use and applicability in the Indian court of law is far beyond admissibility. To validate the proposed framework would be another challenging task for the future researchers so that the technique can be truly useful for assisting the examiners for identity tracing of the unknown suspect.

References

Baayen, R.H., Van Halteren, H., Neijit, A., & Tweedie, F. 2002. An experiment in authorship attribution.   In proceedings of the 6th International conference on the Statistical Analysis of Textual Data, St. Malo, France.

Corney, M., Vel, O. d., Anderson, A. and Mohay, G. 2002.Gender-Preferential Text Mining of Email Discourse. In proceedings of the 18th Annual Computer Security Application Conference, LA, NV

Douglas et al. "Criminal Profiling from Crime Scene Analysis".*Behavioural sciences & the law* 4.4(1986): 401-421.

Koppel et al. "Automatically categorizing written text by author gender". *Literary and linguistic computing* 17.4(2002): 401-412.

Mosteller, F et al. 1964. Interference and disputer authorship: The Federalist. Reading, MA: Addison – Wesley

Stamatatos et al. "Computer based authorship attribution without lexical measures". *Computers and the humanities* 35.2(2001): 193-214.

Yule, G.U. "On sentence as a statistical characteristic of style in prose". *Biometrica* 30 (1938): 363-390.

A Framework for Authorship Identification of Questioned Documents: Forensic and Linguistic Convergence by Sresha Yadav and Smita Jha

Zheng, R., Li, J., Chen, H., and Huang, Z. "A Framework for Authorship Identification of Online Messages: Writing- Style Features and Classification Techniques". *The American society for information science and technology,* 57.3(2006): 378-393.