A Review of Machine Learning based Plagiarism Detection Approaches

Shabbir Attique[1], Warish Hassan[2]

University of Wah,Pakistan [1,2]

ABSTRACT

Author attribution is a subfield of Automatic Language Processing (NLP) which involves identifying the most likely author of a text among a set of candidates. Thus, it can be seen as a text classification problem. This task is part of the forensic sciences, the detection of plagiarism, even the identification of cybercriminals. This is a interdisciplinary field with an intersection of machine learning and Stylometry. Most of the previous works focus on texts while current trends in information technology encourage the use of shorter and more informal texts. Thus, this contribution aims to study the question of author attribution on short messages.

Keywords: Machine Learning, Plagiarism

I. INTRODUCTION

To date, no linguistic analysis has been carried out on the texts submitted to technicians working for the French courts. Author attribution by linguistic analysis would therefore constitute a new field of analysis in the field of Documents and could To do this, she Author attribution using deep learning methods must demonstrate its ability to respond to the issues specific to the environment in which forensic science operates, in order to be usable. Document analysis specialists in the field of forensics analyze traces linked to documents or other assimilated objects. These traces come, for example, from secure documents (passports, drive...), anonymous letters (handwritten, typed or typed), contracts, checks, tags, wills or even administrative forms. Exams may relate to the composition of the paper, the search for non-visible mechanical traces with the naked eye (usually handwritten

notes or marks from the drive rollers of a printer), the determination of printing techniques, the analysis and differentiation of inks and the comparison of handwritings. The delimitation of the working area for the linguist is an essential gateway to understanding the possible fields of action of a discipline still under construction, and respond to the challenges of forensic scientists -The purpose of this article is to report on the advances obtained during the realization of a research project carried out on the textual analysis of writings and the attribution of authors, in order to bring a new axis of research concerning texts analyzed in the field of handwriting comparison in forensic science. With the developments NLP, in particular through deep learning thanks to neural networks, new perspectives are to be considered to move forward in this area. Author attribution is a subfield of NLP that involves identifying the most likely author of a text among a set of candidates. Thus, it can be seen as a text classification problem, including supervised learning. It is a cross between stylometry and machine learning lead to the establishment of a new forensic specialty. Furthermore, contemporary incarnations of technology, such as serverless computing, enable the autonomous deployment of innovative energy-use patterns. The customizable virtual machine based on containers will boost cloud use and offer low latency for the database environment.

II. Related Work

Stylometrics is a branch of computational linguistics that studies literary style using quantitative methods. It supposes that an author unconsciously leaves characteristics in his text that can lead to his identification[1]. Characteristics Stylometrics used in Author Attribution can be separated into different levels analysis (lexical, syntactic, semantic, structural and application specific). These characteristics can be: the number of words, sentences; the length of words, sentences; the frequency of tool words, form words, n-grams of words, n-grams of characters;[2, 3] frequency of parts of speech; collocations (the frequency of bigrams parts of speech); the number of hapax legomena (words appearing only once); etc

Jan et-al presented a benchmark neural paraphrase detection[4], in this research they show the paraphrased text maintains the semantics of the original source. In [5] the author contributed in machine paraphrased plagiarism. There are many other contributions in language model transforming with machine-paraphrased plagiarism[6].

III. Data and methods

The data used in our protocol comes from the corpus of tweets from the. It is true that political tweets are not necessarily representative tweets from the general public. Nevertheless, we find that these types of tweets provide us still the proper characteristics so that we can see the potentiality of this research topic .We chose this data because it is a corpus well-known exploratory test of various works that reliably tests networks. The 2017 corpus includes a total of 42,923 tweets from 11 candidates produced during 2017 elections (see Figure 1). This corpus is then divided into 3 sub-corpus with a 60-20-20 ratio which allows to train, estimate and evaluate the model.

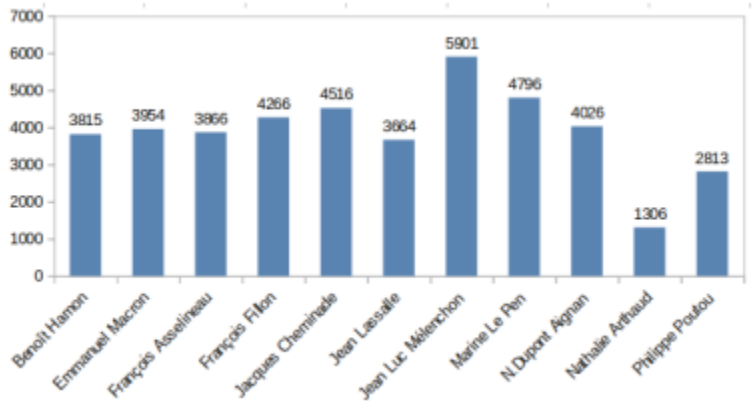

Figure.1 Distribution of tweets by candidate.

IV. CONCLUSION

Companies are producing a wide range of resources by utilizing IoT data and computational software. These tools offer information mining solutions through the application of statistical modeling, prediction, and classification technologies. IoT changes how politicians make decisions. With the development of IoT and related technologies like cloud computing, data sources may be removed from a variety of areas. Current systems would benefit from the emergence of the IoT and AI. Combining computerization and comprehensive analysis, development's benefits are harvested while producing enormous economic gain. The potential to use IoT and artificial intelligence now seem to be greater. The most difficult issues in CC are examined in this study as security vulnerabilities.

References:

[1] G. K. Savova *et al.*, "Use of Natural Language Processing to Extract Clinical Cancer Phenotypes from Electronic Medical RecordsNatural Language Processing for Cancer Phenotypes from EMRs," *Cancer research,* vol. 79, no. 21, pp. 5463-5470, 2019.

[2] I. R. a. Kelley, "Data management in dynamic distributed computing environments," Thesis (Ph.D.), Cardiff University, 2012. [Online]. Available: http://orca.cf.ac.uk/44477/

[3] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional gan," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2970-2979.

[4] J. P. Wahle, T. Ruas, N. Meuschke, and B. Gipp, "Are neural language models good plagiarists? A benchmark for neural paraphrase detection," in *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2021: IEEE, pp. 226-229.

[5] J. P. Wahle, T. Ruas, T. Foltýnek, N. Meuschke, and B. Gipp, "Identifying machine-paraphrased plagiarism," in *Information for a Better World: Shaping the Global Future: 17th International Conference, iConference 2022, Virtual Event, February 28–March 4, 2022, Proceedings, Part I*, 2022: Springer, pp. 393-413.

[6] J. P. Wahle, T. Ruas, F. Kirstein, and B. Gipp, "How Large Language Models are Transforming Machine-Paraphrased Plagiarism," *arXiv preprint arXiv:2210.03568,* 2022.