Unsupervised method for user identification in social media

Haedong Kim
The Pennsylvania State University
Department of Industrial and Manufacturing Engineering
State College, PA 16802, USA
huk344@psu.edu

Conrad S. Tucker
The Pennsylvania State University
Department of Industrial and Manufacturing Engineering
State College, PA 16802, USA
ctucker4@psu.edu

***Abstract*—The abstract goes here.**

INTRODUCTION

Author identification has attracted intense attention because of its broad applications, for example, in finding the original author of papers or plagiarism (Coulthard 2004), computer forensics (De Vel et al. 2001), and fake news detection (Shu et al. 2017). The analysis of authorship is based on the assumption that individuals have distinctive writing style so that it allows us to discriminate between one and other (Swain, Mishra, and Sindhu 2017). This kind of research is originated from the branch of linguistics called stylometry (El and Kassou 2014).

Although the history of author identification goes back to more than a century with the research on Shakespear's plays (Mendenhall 1887), we focus on the online text in this study. More specifically, we are interested in social media textual data like Tweeter tweets. In the following, the special case of author identification, clarifying social media users, will be called *user identification*.

Social media data causes new challenges in author identification because of its distinctive characteristics compared to traditional writings such as literary works, and articles. First, the length of text in social media is usually short (Okuno, Asai, and Yamana 2014). Postings in social media tend to be terse than traditional writings, or even other online text such as blogs, and emails. It is obvious that the longer text is, the more linguistic information needed to clarify the author is there (Brocardo et al. 2013, Okuno, Asai, and Yamana (2014)). Second, J. S. Li et al. (2017) reports that in many cases the number of postings from some user is not enough to learn the writing style of the author. Third, social media data does not have much reliable meta-data could be used to the user identification. In Meyer and Bernstein (2007), the authors use email meta-data such as email addresses, organization names, phone numbers. In contrast, in social media, some user accounts do not have personal information, or even though there is such information, it is not reliable because users can write down it without verification. Last, social media data is very noisy and unstructured as it includes a lot of improper grammar and spelling, as well as acronyms and newly coined buzzwords (Gundecha and Liu 2012).

In the existing studies, a standard procedure of author identification is first to represent documents as a vector using stylometry features and train classification models on the set of documents embedded in a vector space (**add ref**). The same process applies to the user identification task. In this case, documents are tweets or postings. Stylometry features are a function mapping texts to real numbers, so that allow us to apply operators, such as similarity functions, to the textual data (**add ref**).

Although the existing studies have shown promising results, there are some limitations. First, each study adopts a subset of all possible features. It is crucial given the fact that it has been verified the richer set of features yields the better performance (Zheng et al. 2006, Hurtado, Taweewitchakreeya, and Zhu (2014)). Second, most existing methods are based on supervised learning (i.e., classification models) (J. S. Li et al. 2017)(**add ref**). Supervised learning is not practical for identifying social media users, because training data of texts generated by authentic users is necessary for supervised learning. It implies that only a pre-determined set of users can be identified, and some amount of time of collecting normal texts is required. It limits a range of applications and, as aforementioned, some users do not write many postings.

To tackle these problems, this study has conducted the exhaustive search for finding almost all possible stylometric features and propose an unsupervised user identification method.

The paper is organized as follows. Section 2 reviews the related work to this study. Section 3 describes the proposed methodology.

LITERATURE REVIEW

We review the existing literature on stylometry features and user identification methods in social media. Identification methods are into two categories: supervised and unsupervised approaches.

A. Stylometry features for author identification

Five categories lexical, syntactic, structural, content-specific, idiosyncratic

B. Supervised user identification methods

summarize disadvantages of supervised methods

C. Unsupervised user identification methods

flaws of existing unsupervised methods
summarize noveties of out methods in a table

APPROACH

Proposed method

EXPERIMENT

Experiment with Tweeter data and discuss about the application of author identification in fake news detection.

CONCLUSION

The conclusion goes here.

ACKNOWLEDGMENT

The authors would like to thank...

REFERENCES

Brocardo, Marcelo Luiz, Issa Traore, Sherif Saad, and Isaac Woungang. 2013. "Authorship Verification for Short Messages Using Stylometry." In *Computer, Information and Telecommunication Systems (Cits), 2013 International Conference on*, 1–6. IEEE.

Coulthard, Malcolm. 2004. "Author Identification, Idiolect, and Linguistic Uniqueness." *Applied Linguistics* 25 (4). Oxford University Press: 431–47.

De Vel, Olivier, Alison Anderson, Malcolm Corney, and George Mohay. 2001. "Mining E-Mail Content for Author Identification Forensics." *ACM Sigmod Record* 30 (4). ACM: 55–64.

El, Sara El Manar, and Ismail Kassou. 2014. "Authorship Analysis Studies: A Survey." *International Journal of Computer Applications* 86 (12). Foundation of Computer Science.

Gundecha, Pritam, and Huan Liu. 2012. "Mining Social Media: A Brief Introduction." In *New Directions in Informatics, Optimization, Logistics, and Production*, 1–17. Informs.

Hurtado, Jose, Napat Taweewitchakreeya, and Xingquan Zhu. 2014. "Who Wrote This Paper? Learning for Authorship de-Identification Using Stylometric Featuress." In *2014 Ieee International Conference on Information Reuse and Integration (Iri)*, 859–62. IEEE.

Li, Jenny S, Li-Chiou Chen, John V Monaco, Pranjal Singh, and Charles C Tappert. 2017. "A Comparison of Classifiers and Features for Authorship Authentication of Social Networking Messages." *Concurrency and Computation: Practice and Experience* 29 (14). Wiley Online Library: e3918.

Mendenhall, Thomas Corwin. 1887. "The Characteristic Curves of Composition." *Science* 9 (214). JSTOR: 237–49.

Meyer, David, and Steven Miller Bernstein. 2007. "Meta-Content Analysis and Annotation of Email and Other Electronic Documents." Google Patents.

Okuno, Syunya, Hiroki Asai, and Hayato Yamana. 2014. "A Challenge of Authorship Identification for Ten-Thousand-Scale Microblog Users." In *Big Data (Big Data), 2014 Ieee International Conference on*, 52–54. IEEE.

Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. "Fake News Detection on Social Media: A Data Mining Perspective." *ACM SIGKDD Explorations Newsletter* 19 (1). ACM: 22–36.

Swain, Siddharth, Gaurav Mishra, and C Sindhu. 2017. "Recent Approaches on Authorship Attribution Techniques—An Overview." In *Electronics, Communication and Aerospace Technology (Iceca), 2017 International Conference of*, 1:557–66. IEEE.

Zheng, Rong, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. "A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques." *Journal of the American Society for Information Science and Technology* 57 (3). Wiley Online Library: 378–93.