Digital Investigation 16 (2016) S33–S43

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

DFRWS 2016 Europe — Proceedings of the Third Annual DFRWS Europe

Authorship verification for different languages, genres and topics

Oren Halvani*, Christian Winter, Anika Pflug

Fraunhofer Institute for Secure Information Technology SIT, Rheinstr. 75, 64295 Darmstadt, Germany [1]

Keywords:
Digital text forensics
Intrinsic authorship verification
One-class-classification
Cross-genre
Cross-topic

ABSTRACT

Authorship verification is a branch of forensic authorship analysis addressing the following task: Given a number of sample documents of an author $\mathscr{A}$ and a document allegedly written by $\mathscr{A}$, the task is to decide whether the author of the latter document is truly $\mathscr{A}$ or not. We present a scalable authorship verification method that copes with this problem across different languages, genres and topics. The central concept of our method is a model, which is trained with Dutch, English, Greek, Spanish and German text documents. The model sets for each language specific parameters and a threshold that accepts or rejects the alleged author as $\mathscr{A}$. The proposed method offers a wide range of benefits, e.g., a universal (static) threshold for each language and scalability regarding almost any involved component (classification function, ensemble strategy, features, etc.). Furthermore, the method benefits from low runtime due to the fact that no natural language processing techniques nor other computationally-intensive methods are involved. In our experiments, we applied the method on 28 test corpora including 4525 verification cases across 16 genres and a huge number of mixed topics, where we achieved competitive results (75% median accuracy). With these results we were able to outperform two state-of-the-art baselines, given the same training and test corpora.

Introduction

Forensic authorship analysis forms a branch of digital forensics with many application scenarios. There are a lot of real-world cases where a document has a certain alleged author, but the authorship is disputed by another party. Examples are university theses suspected that they have been written by a ghostwriter (Mothe et al., 2015), fictive insurance claims invented by a field agent of an insurance company, a forged last will or sock puppet detection (also known as multiple account detection) (Afroz et al., 2014). The underlying forensic task is *authorship verification*. Authorship verification can also be applied for deciding whether two documents originate from the same author, e.g., for ascertaining whether two illegal online services, like illegal drug stores or illegal file distribution platforms, are operated by the same person. Due to the importance of authorship verification for forensic trials, this paper focuses on this task.

Another kind of forensic authorship analysis is *authorship profiling*, which determines author specific attributes such as gender, age or regional and social background of an unknown author, e.g., a blackmailer. A third variant is *authorship attribution* (also known as authorship identification), which tries to determine the true author in cases where several people are suspected as author of a document like a threatening letter, a terroristic proclamation, or a book published under a pseudonym.

* Corresponding author.
E-mail addresses: Oren.Halvani@SIT.Fraunhofer.de (O. Halvani), Christian.Winter@SIT.Fraunhofer.de (C. Winter), Anika.Pflug@SIT.Fraunhofer.de (A. Pflug).

[1] www.SIT.Fraunhofer.de.

http://dx.doi.org/10.1016/j.diin.2016.01.006

Authorship attribution (AA) deals with the problem of attributing a given anonymous text to one author, given a set of candidate authors for whom text samples of undisputed authorship are available (Stamatatos, 2009). The set of all authors with their corresponding text samples is usually referred to as *reference set*, which is analyzed regarding the writing style of the candidates in order to compare it to the writing style of the anonymous author. From this, a decision about the true author is made. If it is guaranteed that the reference set contains the true author, the AA task is considered to be a closed-set and otherwise an open-set problem. The majority of existing studies focuses on closed-set problems (Stolerman et al., 2014), which are known to be easier to solve than open-set problems (Potha and Stamatatos, 2014).

Authorship verification (AV) addresses the problem to determine whether a given text was written by a certain author $\mathscr{A}$ or not (Stamatatos, 2009). The reference set in AV comprises only text samples from the supposed author $\mathscr{A}$, that can be analyzed for distinct features that describe the writing style of $\mathscr{A}$. If it differs significantly from the writing style of the given text, the authorship is considered false, otherwise true. AA can be treated as a sequence of AV problems regarding the given candidate authors (Koppel et al., 2012) and should be solvable if the verification problem is solved (Koppel and Winter, 2014). Conversely, AV represents a specialized task of AA with an open set of candidate authors. In this regard, AV can be understood as a *one-class classification problem* (Stein et al., 2008). This means a text is either classified as part of the given class, i.e. attributed to the supposed author, or classified as an outlier or negative example of the class (Tax, 2001).

If modeled as a one-class classification task, AV is a more difficult problem. This is mostly due to the challenge of determining boundaries between class elements and outliers without negative examples or at least exhaustive and representative positive examples (Koppel and Schler, 2004). In spite of its difficulty, AV is a rewarding topic because it often occurs as a task in practice, and it can also aid in other tasks as, for example, intrinsic plagiarism detection (Stein et al., 2008).

Challenges

AV, as a special case of AA, faces partially the same challenges. Similar to AA in general, some obstacles are, according to Stamatatos (2009):

- Sparse, unequally distributed, non-representative or difficult text sources (e.g., colloquial phrasing, short texts, noise from careless or author-typical mistakes).
- Coping with influences from topic, genre, time period or language of the document.
- Accuracy of involved tools (e.g., natural language processing tools).
- Finding appropriate features for the given task.

AV as a one-class classification problem faces challenges that are common to classification tasks. However, the biggest obstacle in AV is determining a universal threshold for separating genuinely authored texts from negative examples. In comparison, general AA has the advantage of balancing this decision between available positive and negative examples. In theory, negative examples can also be found for AV because all texts which were definitely not written by the supposed author represent outliers. However, due to the large amount of these possible examples, selecting those which are most representative or encompassing is difficult (Koppel and Schler, 2004). Hence, in comparison to AA, it is also more challenging to determine the most important text features for distinguishing authors. The selected features might only work best with the gathered negative samples and fail in other cases.

Furthermore, the decision boundary or threshold has to be adjusted to the application area of the verification method. It might be desirable to have a very liberal or conservative decision. This means the method either identifies a lot of same-author texts while also falsely identifying negative examples, or it only categorizes the most certain texts as same-author.

Extrinsic versus intrinsic methods

According to Stamatatos et al., an AV method can be either *intrinsic* or *extrinsic* (Stamatatos et al., 2014). The authors explain that intrinsic methods rely only on the questioned document $\mathscr{D}_{\mathscr{A}_?}$ and the reference set $\mathbb{D}_{\mathscr{A}}$ of the known author $\mathscr{A}$ for deciding the verification task. In contrast, extrinsic methods make use of additional documents by other authors in order to transform AV from a one-class-classification task to a binary classification task.

Our proposed AV method is an intrinsic method, since we do not use additional texts of other authors for deciding whether, the $\mathscr{D}_{\mathscr{A}_?}$ has been written by $\mathscr{A}$ or not.

Contribution

We propose a new intrinsic AV method that offers a number of contributions. Our method provides a universal threshold per language, used to accept or reject the alleged authorship of a document. Here, *universal* refers to the ability to generalize across different genres and topics of the texts. In addition, the model generated by the method is flexible and can be extended incrementally in order to handle new languages, genres or features. Our method does not involve error-prone natural language processing techniques nor machine learning libraries (e.g., Weka, RapidMiner, Scikit or Shogun), which often encapsulate the classification task as a black box. Furthermore, the method is compact and fully transparent and thus, can be reimplemented easily by the community. Moreover, it features a low computational complexity (on average, few seconds per case) compared to other existing approaches (e.g., listed in Stamatatos et al. (2014)). The evaluation of the proposed method on 28 test corpora distributed over five languages, 16 genres, and mixed topics shows, that the performance is stable in terms of accuracy, even for uncommon AV scenarios such as diploma theses or cooking recipes.