Poster: Assessing the Effectiveness of Countermeasures Against Authorship Recognition

Michael Backes
CISPA, Saarland University
backes@cs.uni-saarland.de

Pascal Berrang
CISPA, Saarland University
berrang@cs.uni-saarland.de

Praveen Manoharan
CISPA, Saarland University
manoharan@cs.uni-saarland.de

***Abstract*—Methods for authorship recognition were originally developed to aid in criminal investigations and attribution of historical texts. Nowadays, however, in an age in which the Internet has become the central platform for day-to-day social interactions and communication, authorship recognition technology can be abused to break the anonymity of users by identifying the authors of user-generated text content. While there have been recent advances in *adversarial stylometry*, which investigates the impact of *obfuscation* and *imitation* on current authorship recognition techniques, no comprehensive model for the assessment of countermeasure effectiveness currently exists.**

In this work, we introduce a novel measure for assessing the context-dependent importance of writing style features in authorship recognition. From this measure, we furthermore derive an additional measure for assessing the effectiveness of authorship-recognition countermeasures by analysing how well these countermeasures reduce the importance of the affected features.

We then utilise these measures to conduct a large-scale evaluation of four semantics-retaining countermeasures and their combinations on a dataset of $923{,}997$ comments collected from the online social network Reddit. We examine the practical impact of these countermeasures on the importance of standard writing style features and explore the outcome of combining several countermeasures at the same time. Moreover, we validate our approach by applying it to the *Extended-Brennan-Greenstadt Adversarial Corpus* and comparing our results with previous research.

I. Motivation

During the last two decades, the Internet evolved from a simple communication network to a global multimedia platform which is part of our everyday life. On this platform, billions of users actively share data, even revealing personal information, without considering the consequences of the easy accessability and the permanent nature of their disseminated data. The detrimental consequences range from personalized advertisements and the sale of personal information up to threats concerning personal safety.

An intuitive, commonly pursued approach to separate sensitive information from one's personal identity, and thus protect one's privacy, would be to disseminate sensitive information only through anonymous or pseudonymous profiles, with the intention of decoupling a user's real-life identity from sensitive information posted under pseudonymous accounts. As literature has shown, however, this approach is not really effective since different profiles are typically linkable using common characteristics [1], [2], [3], [4], [5]. In particular, for user-generated text content, the writing style of a user is often unique across different profiles, and can thereby be used to attribute text content to its corresponding (seemingly anonymous) author [6]. Current research has shown that this profile linkage can even be conducted at Internet scale [7].

Recent work on *adversarial stylometry* [8], [9] has tried to reduce the likelihood of correctly linking corresponding profiles by investigating the impact of *obfuscation* and *imitation* of text passages on current authorship recognition techniques. These works have mostly focused on the development of manual and semi-automated countermeasures in order to circumvent stylometry. However, none of these works is capable of assessing the actual effectiveness of these countermeasures. Hence, these works do not provide any insights on which countermeasures are particularly well-suited for a given context in which a certain text should be published. The absence of such results is, in particular, due to the lack of a rigorous model for assessing the effectiveness of various types of countermeasures on the identifiability of authors, which currently does not exist. In addition, developing a model of this kind might help in identifying the major challenges that research needs to overcome in order to provide fully-automated assistance for authorship obfuscation.

II. Novel Measure for Feature Importance

In this work, we present a novel, classifier-independent measure for assessing the importance of stylometric features for the identifiability of authors. We base this assessment on the privacy model introduced by Backes *et al.* [10], which provides a generic data model to cope with heterogeneous information using statistical models. We adapt and extend these statistical models to fit our use case to authorship recognition, effectively defining a model for writing style that allows us to capture a comprehensive list of stylometric features, as introduced by Abbasi and Chen [3]. Overall, we develop a model of the authorship recognition problem that allows us to formally reason about authorship recognition in the open setting of the Internet.

We then derive how we can identify important stylometric features that significantly contribute to the identification of the correct author from the context in which text is published by using these writing-style models. We employ standard regression techniques to determine the weights of each type of stylometric feature, which then correspond to their importance.

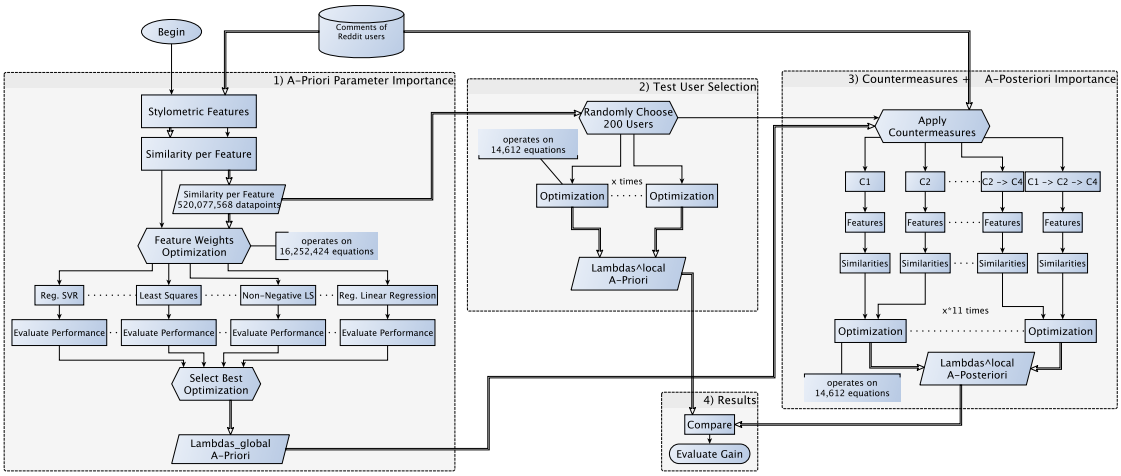

Figure 1: This flowchart represents our actual set-up.

III. Assessing the Effectiveness of Countermeasures

From the features' importance assessment we then further derive the *gain* measure for the effectiveness of countermeasures against authorship identification by measuring how well they reduce the importance of stylometric features towards a target value. This target value depends on the actual importance values and identifies the desired state that minimizes the success of authorship recognition techniques.

Definition 1 (Gain): Let $\lambda_1, \ldots, \lambda_n$ be the features' importance values, $\varphi_1, \ldots, \varphi_n$ the corresponding target values and $\lambda'_1, \ldots, \lambda'_n$ the corresponding values after the application of a countermeasure. Then the *gain* provided by the countermeasure is defined as $\text{gain} = \sum_i |\lambda_i - \varphi_i| - |\lambda'_i - \varphi_i|$.

IV. Evaluation and Validation

We apply this measure to assess the effectiveness of four, fully automated countermeasures, namely synonym substitution, spell checking, special character modification and adding/removing misspellings. Moreover, we group them into two categories: informed countermeasures, which choose their action based on a local optimization, and uninformed countermeasures, which always choose a fixed or random action without involving any further knowledge.

As shown in Figure 1, we follow a general and comprehensive methodology that structures the evaluation process and is easily extensible for future evaluation.

First, we perform our experiments on a dataset of 923,997 comments by 3439 users collected from the online social network Reddit. In particular, we apply the countermeasures individually, and also in combination, to the comments of test users and evaluate the gain provided on average. The results of this evaluation are already available and suggest a surprising improvement on the importance values in case of the spell checker. We also show that our gain measure correlates with authorship recognition precision.

Then, we also apply our methodology on the *Extended-Brennan-Greenstadt Adversarial Corpus* [8], [9] and validate our approach by giving a comprehensive comparison to existing authorship recognition techniques.

As part of our work in progress, we are currently facing the following challenges: the actual interpretation of possibly negative features' importance values – due to non-negative regression techniques – and tracing back the gains provided by a countermeasure to its actual implementation in order to allow for a more detailled interpretation of their gains.

References

[1] T. C. Mendenhall, "The characteristic curves of composition," *Science*, pp. 237–249, 1887.

[2] O. Uzuner and B. Katz, "A Comparative Study of Language Models for Book and Author Recognition," in *Proc. of the 2nd International Joint Conference on Natural Language Processing (IJCNLP)*, 2005, pp. 969–980.

[3] A. Abbasi and H. Chen, "Writeprints: A Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 2, pp. 1–29, 2008.

[4] M. Koppel, J. Schler, and S. Argamon, "Computational Methods in Authorship Attribution," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 1, pp. 9–26, 2009.

[5] M. B. Malyutov, "Authorship Attribution of Texts: A Review," in *General Theory of Information Transfer and Combinatorics*, 2006, pp. 362–380.

[6] S. Afroz, A. C. Islam, A. Stolerman, R. Greenstadt, and D. McCoy, "Doppelgänger finder: Taking stylometry to the underground," in *Proc. of the 2014 IEEE Symposium on Security and Privacy*, 2014, pp. 212–226.

[7] A. Narayanan, H. Paskov, N. Z. Gong, J. Bethencourt, E. Stefanov, E. C. R. Shin, and D. Song, "On the feasibility of internet-scale author identification," in *Proc. of the 2012 IEEE Symposium on Security and Privacy (S& P)*, 2012, pp. 300–314.

[8] M. R. Brennan and R. Greenstadt, "Practical Attacks Against Authorship Recognition Techniques," in *Proc. of the 21st Annual Conference on Innovative Applications of Artificial Intelligence (IAAI)*, 2009.

[9] M. R. Brennan, S. Afroz, and R. Greenstadt, "Adversarial Stylometry: Circumventing Authorship Recognition to Preserve Privacy and Anonymity," *ACM Transactions on Information and System Security (TISSEC)*, vol. 15, no. 3, pp. 12:1–12:22, 2012.

[10] M. Backes, P. Berrang, and P. Manoharan, "How well do you blend into the crowd? - d–convergence: a novel paradigm for reasoning about privacy in the age of Big-Data," http://arxiv.org/abs/1502.03346, 2015.