Authorship Attribution for Social Media Forensics

Sudeep Shrestha, Shubham SenGupta, Prema Ale, Vaishnavi V Naik

Department Of Computer Science and Engineering, Brindavan College of Engineering, Bangalore-063, India.

***Abstract:** The veil of anonymity provided by smart phones with pre-paid SIM cards, public Wi-Fi hotspots, and distributed networks like Tor has drastically complicated the task of identifying users of social media during forensic investigations. In some cases, the text of a single posted message will be the only clue to an author's identity. How can we accurately predict who that author might be when the message may never exceed 140 characters on a service like Twitter? For the past 50 years, linguists, computer scientists and scholars of the humanities have been jointly developing automated methods to identify authors based on the style of their writing. All authors possess peculiarities of habit that influence the form and content of their written works. These characteristics can often be quantified and measured using machine learning algorithms. In this article, we provide a comprehensive review of the methods of authorship attribution that can be applied to the problem of social media forensics. Further, we examine emerging supervised learning based methods that are effective for small sample sizes, and provide step-by-step explanations for several scalable approaches as instructional case studies for newcomers to the field.*

***Keywords:** Authorship attribution; Computational linguistics; Forensics; Machine learning; Social media*

I. INTRODUCTION

This paper lies under the domain of data mining. There has been a huge increase in the amount of data being stored in databases as well as the number of database applications in business and the scientific domain. This explosion in the amount of electronically stored data was accelerated by the success of the relational model for storing data and the development and maturing of data retrieval and manipulation technologies. While technology for storing the data developed fast to keep up with the demand, little stress was paid to developing software for analyzing the data until recently when companies realized that hidden within these masses of data was a resource that was being ignored.

A. Purpose of Data Mining

While the main concern of database technologists was to find efficient ways of storing, retrieving and manipulating data, the main concern of the machine learning community was to develop techniques for learning knowledge from data. Data Mining can be considered to be an inter-disciplinary field involving concepts from Machine Learning, Database Technology, Statistics, Mathematics, Clustering and Visualization among others.

B. Existing System

Earlier some researchers joined the messages in a single document. It is assumed that the source of a text is either one author (or possibly several) out of a known set, or an author that is unknown to investigators. Automated approaches to authorship attribution via the methods of statistical pattern recognition have been around for decades. Social media messages from are very short and therefore a smaller set of stylometric features is present in each one.

C. Proposed System

The proposed model in this paper initially loads the data from the dataset that has been posted in the various social media. A sentimental analysis is performed on these data. This analysis includes preprocessing of the data. The data are preprocessed in such a way that key words in the data are fetched and the data is cleaned. They are further classified using bag of words model. These authors are given pseudonyms. After this those messages are given as input to the classifier as a test data set. Upon classification they are classified as good or bad. If found bad, they are matched with the known authors. Such authors are revoked and the messages are deleted.

II. METHODS OF AUTHORSHIP ATTRIBUTION

A. General Stylometric Features for Forensic Authorship Attribution:

At the most basic level, the words of a text are useful features in and of themselves for authorship attribution. However, all words cannot simply be treated as features. Thus, it is common to discard the function words, those words that occur most frequently but carry little if any semantic meaning (Table I) to isolate a more stable signal. However, function words can be useful for attribution in some cases. For instance, function words can be coupled with the most frequent punctuation, or other stable features, becoming more flexible and discriminative for use in a bag-of-words representation [1] that disregards grammar, but preserves the underlying statistics of Language. TableI: Function words are a very basic, but sometimes useful feature for authorship attribution in all contexts. They can be particularly effective for social media analysis, because they tend to be the words that occur most frequently. Thus the probability of occurrence in even small samples like tweets is high.

B. General Classifiers for Forensic Authorship Attribution

Once a feature set has been chosen, the next step is to select a classification method. Authors are then classified via data clustering. Multivariate analysis achieved some measure of success, and it quickly became well-established for authorship attribution. However, the presence of some labelled data often improves results dramatically. Accordingly, supervised approaches now dominate the field.

C. Authorship Attribution for Small Samples of Text:

What do we do when we do not have that luxury? A tweet, for instance, is a mere 140-characters long, and does not yield a large amount of information at the word-level, or from its syntactic structure. Even before the advent of social media, researchers had been investigating this problem in the context of forensic stylometry in e-mails, where short form writing is the norm. Some of the strategies we discussed above, such as similarity measures and SVM apply directly, but better performance can be achieved with features and classification approaches custom-tailored for attribution problems with small samples of texts.

D. Authorship Attribution Specifically for Social Media:

A growing body of work has attempted to mine and analyse actual online postings. Abbasi and Chen produced a body of work that included an analysis of eBay comments and online forum posts.

E. The Threat of Counter-Attribution:

Naturally, in authorship attribution, there exists some element of the ―offense and defence ‖ dynamic present in the broader world of computer security. Counter-attribution techniques, where there is an intentional act of changing one's writing style, have emerged to thwart authorship attribution system.

III. Walk -Through Of Authorship Attribution Techniques For Social Media Forensics:

A. General Framework for Authorship Attribution

Refer Fig. 1 for general framework

B. Source Data from Twitter:

Two preliminary steps are required before feature extraction and classification: data extraction and text pre-processing.

Data Extraction: Since the ultimate forensic goal is authorship attribution, all the retweets, which are messages retransmitted from other users, should be removed. This is done by removing all tweets marked by the Twitter API with a specific retweet flag, as well as tweets containing the meta tag RT [10], [11]. Our focus is on English-language tweets, thus non-English tweets can be removed using the python library guess-language [12], which itself uses the spell-checking library penchant [13] to build an accurate prediction of the language of a text consisting of three or more words. For this reason, we recommend removing all messages that contain only one or two words. This is not done to the detriment of accuracy — these short messages do not provide meaningful information about the author and end up introducing noise into the classification task [11].

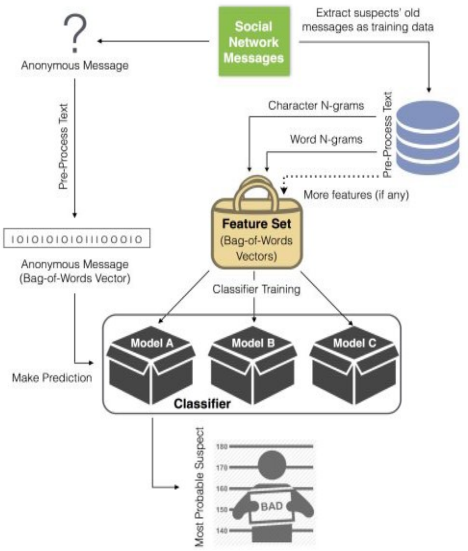

Fig 1. General Framework

Example;

Before pre-processing:

Tweet 1: ―Do not forget my bday is on 03/27 #iwantgifts‖

Tweet 2: ―@maria I will be sleeping @00:00AM‖

Tweet 3: ―Check out this amazing website: http://www.ieee.org‖

After pre-processing:

Tweet 1: ―Do not forget my bday is on DAT TAG |

Tweet 2: ―REF I will be sleeping @TIM‖

Tweet 3: ―Check out this amazing website: URL‖

C. Bag-of-Words Model:

The bag-of-words is a classic model in natural language processing. For natural language processing tasks such as information retrieval and sentiment analysis, it has been suggested that function words should be removed [14].

IV. Conclusion

The project thus makes the user to be registered before uploading or posting any message. The pattern of the author is watched accordingly whenever he enters into the website and attempts for a search or post any messages. The words or messages that are posted by the user is extracted by using lexical analysis and the exact meaning of the word is found and fetched. If that word is found to be the bag of words then that particular user is warned and he cannot post further messages in the website. This kind of processing the messages will be helpful in finding the bad users or the users spreading the messages unwantedly. Since the particular user is been blocked and stopped from being sending the messages the rumors can be easily stopped.

REFERENCES

[1] D. Jurafsky. Speech & language processing, second edition. Prentice Hall, 2008.

[2] H. Baayen, H. van Halteren, A. Neijt, and F. Tweedie. An experiment in authorship attribution. In Journal of Textual Data Statistical Analysis, pages 29–37. Citeseer, 2002.

[3] H. Baayen, H. Van Halteren, and F. Tweedie. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. Literary and Linguistic Computing, 11(3):121–132, 1996.

[4] J. N. G. Binongo. Who wrote the 15th book of Oz? an application of multivariate analysis to authorship attribution. Chance, 16(2):9–17, 2003.

[5] M. L. Brocardo, I. Traore, S. Saad, and I. Woungang. Authorship verification for short messages using stylometry. In Intl. Conference on Computer, Information and Telecommunication Systems, pages 1–6, 2013.

[6] M. Koppel, J. Schler, and S. Argamon. Authorship attribution in the wild. Language Resources and Evaluation, 45(1):83–94, 2011.

[7] M. Koppel and Y. Winter. Determining if two documents are written by the same author. Journal of the Association for Information Science and Technology, 65(1):178–187, 2014.

[8] T. Qian, B. Liu, L. Chen, and Z. Peng. Tri-training for authorship attribution with limited training data. In Annual Meeting of the Association for Computational Linguistics, pages 345–351, 2014.

[9] A. Abbasi and H. Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Transactions on Information Systems, 26(2):7, 2008.

[10] R. Layton, P. Watters, and R. Dazeley. Authorship attribution for Twitter in 140 characters or less. In Cybercrime and Trustworthy Computing Workshop, pages 1–8, 2010.

[11] R. Schwartz, O. Tsur, A. Rappoport, and M. Koppel. Authorship attribution of micro-messages. In Conference on Empirical Methods on Natural Language Processing, pages 1880–1891. ACL, 2013..

[12] Spirit. Guess Language: Guess the natural language of a text, 2014. Accessed on July 1, 2015 via https://bitbucket.org/spirit/guess_language.

[13] R. Kelly. pyenchant: a spellchecking library for python, 2014. https://pythonhosted.org/pyenchant/.

[14] C. D. Manning and H. Sch¨utze. Foundations of Statistical Natural Language Processing. MIT Press, 1999.