International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 05 | May-2018 www.irjet.net p-ISSN: 2395-0072

AUTHORSHIP ATTRIBUTION USING STYLOMETRY

Sneha H.P[1], Preethi Nadiger[2], Archana M.R[3], Deekshitha.J[4],Jagruthi H[5]

[1,2,3,4] *VTU & Information Science and Engineering BNMIT*
[5]*Assistant Professor, Dept. of Information Science and Engineering, BNMIT, Karnataka, India*

Abstract - *Distributed networks like Tor has made hard the task of identifying users of social media during forensic investigations. In certain fields, the data of a single posted information will be the only clue to an users identity. It may be hard to identify the user in such cases. Since 50 years people are developing automated methods to identify users based on writing styles. All authors have certain habit that influence the form and content of their written works. Machine learning Algorithms can identify such qualities that are similar. Certain methods of authorship attribution that can be applied to the problem of social media forensics are discussed. Examinination 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. There is an argument that there is a need in forensics to new authorship attribution algorithms that can exploit context, can process multimodal data, and are tolerant to incomplete knowledge of the space of all possible authors at training time.*

1. INTRODUCTION

IT is well known that the real lives of Internet users sometimes turn out to be entirely different from who they appear to be online, the nature and consequence of this phenomenon are changing. A recent expose in the New York Times Magazine documented the case of a Russian media agency that allegedly executed organized information campaigns on social media using pseudonyms and virtual identities. It is assumed that some of these campaigns were state sponsored. With an office full of media professionals, the agency achieved success in promoting false news events and influencing public opinion on politics, and was even able to deceive the journalist covering the story for the Times. On the Internet, this practice is known as “trolling” a favorite pastime of bored adolescents, pundits, and unscrupulous social media coordinators. The organization and scale of these trolling campaigns, however, suggests that the practice has moved into a new phase, whereby corporations and governments seek to control the discourse surrounding popular events (both real and imagined) on social media. This poses a legal and security dilemma on multiple fronts. If the underlying identities of the Russian media agency’s employees could be automatically determined, content originating from them could subsequently be flagged and monitored or blocked. However, the Times discovered that the agency always routed its Internet traffic through proxy servers, thus rendering useless the easy path to doing so via the originating IP addresses.

Forensic authorship attribution is the process of inferring something about the characteristics of an author from the form and content of their writing present in a collection of evidence. The emergence of social media as a primary mode of communication has challenged the traditional assumption that a forensic investigation will have access to long form writing (i.e., letters and emails). In this article, we frame the problem as a computational pipeline, in which features are extracted from very small samples of text, and scalable supervised learning is deployed to train author-specific models and make predictions about unknown samples The goal of this system is to prevent the user from posting sensitive data in social network and also gives a statistics to the Admin about the user who is frequently trying to post sensitive messages in social network. Sometimes people post offensive messages on a particular wall which may cause a serious problem to user's reputation. To avoid such kind of serious problem we can apply Information Filtering (IF) technique. This system uses use N-Gram technique for content-based filtering and Weight-age concept for policy-based filtering method. With the help of these concept this system will detect whether the post contains sensitive data or not.

2. EXISTING SYSTEM

Web-based services are used to extract the significant information from large quantity of data respectively. For example Facebook is the most popular social networking site in which millions of people have opened their user account. Facebook provides all type of services like adding friends, recommending friends, sharing of images, audio and video etc. But Facebook also provides facility to user to post the message on wall. So, there is possibility that posted message could be vulgar or offensive one. Which may cause serious problems like harassing or blackmailing can also happen, it means instead of all those advantages there are some disadvantages with Social networking sites.

Disadvantages of the Existing System

- User can post any kind of message which will create conflict in the society.
- User can catalyst some small issues and make it big problem.

3. IMPLEMENTATION

The goal of this system is to prevent the user from posting sensitive data in social network and also gives a statistics to the Admin about the user who is frequently trying to post

sensitive messages in social network. The fig 1 illustrates the post offensive messages on a particular wall which may cause a serious problem to user's reputation. To avoid such kind of serious problem we can apply Information Filtering (IF) technique. This system uses use N-Gram technique for content-based filtering and Weight-age concept for policy-based filtering method. With the help of these concept this system will detect whether the post contains sensitive data or not.

Identifying sensitive post and block the post, which avoid unnecessary conflicts in the society. Gives Statistic Report to admin regarding which user is trying to post sensitive post frequently. The basic strategy we will look at relies on a set of features capturing patterns extracted from the original texts in a bag -of-words model dynamically created for a set of users. When creating a bag-of-words model, can consider one model for a set of authors or one model per author. A dynamic model for each author could allow for a more fine-grained stylometry evaluation, while a general bag of the most frequent patterns comprising many authors at the same time may overlook some discriminative features of particular authors, as those features may not be strong enough to appear globally.

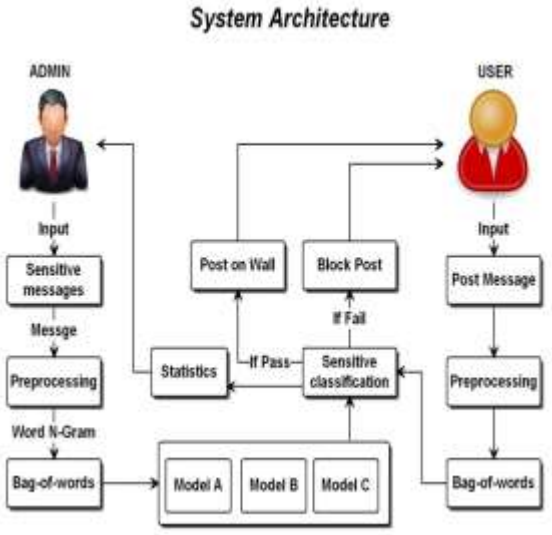

Fig -1: **Design of system architecture**

However, as the number of authors increases, creating a dynamic model for each investigated author is much more time consuming. In both cases, the bag-of-words model works as a projection space in which we aim to highlight similar patterns of a particular author or set of authors, while decreasing other patterns not in accordance with the overall style for those authors of interest. For computational efficiency, we will consider the case of a model per set of authors.

3.1 Feasibility Study

Feasibility is the determination of whether or not a project is worth doing. The process followed in making this determination is called feasibility Study. This type of study if a project can and should be taken. In the conduct of the feasibility study, the analyst will usually consider seven distinct, but inter-related types of feasibility.

3.1.1 Technical Feasibility

This is considered with specifying equipment and software that will successful satisfy the user requirement the technical needs of the system may vary considerably but might include The facility to produce outputs in a given time. Response time under certain conditions. Ability to process a certain column of transaction at a particular speed.

3.1.2 Economic Feasibility

Economic analysis is the most frequently used technique for evaluating the effectiveness of a proposed system. More commonly known as cost / benefit analysis. The procedure is to determine the benefits and savings are expected form a proposed system and a compare them with costs. It benefits outweigh costs; a decision is taken to design and implement the system will have to be made if it is to have a chance of being approved. There is an ongoing effort that improves in accuracy at each phase of the system life cycle.

4. ALGORITHM

4.1 Preprocessing

To our knowledge, no publicly available data set exists for authorship attribution applied to social media forensics. Moreover, the restrictive terms of use put in place by the major social networks prohibit the dissemination of such data sets. Thus, data from the existing research described are largely inaccessible to us. In response to this, we created our own large-scale data set that was designed with algorithm scalability evaluations in mind. The set was constructed by searching Twitter for the English language function words present in Appendix A , yielding results from English speaking public users17. These results were used to build a list of public users from which we could extract tweets by using the Twitter API. We collected ten million tweets from 10,000 authors18 over the course of six months in 2014. Each tweet is at most 140-character long and includes hash tags, user references and links. Although we could not use data from other researchers due to the restrictions placed on us by Twitter's terms of service, the data set was created with the same methods used by other authors. While we cannot release the actual messages, we will release all of the features derived from them after this paper is published in

an effort to provide the community with a standardized resource for evaluation. Pre-processing of each tweet includes removing all non English tweets, tweets with less than four words, and tweets marked as retweets or any tweet containing the meta tag . As discussed previously, for most of the methods we replace numbers, URLs, dates and timestamps by the meta tags NUM, URL, DAT, and TIM,

respectively. Moreover, the hash tags and user references were replaced, since they enrich the feature set for authorship attribution in such a way that makes the task artificially easier yet ultimately unreliable For PPM-5 and SCAP, each digit of a token that does not include letters is instead replaced by the same symbol. The data set was partitioned into training and test sets via k-fold cross validation. each experiment was repeated 10 times and the authors considered in each fold are chosen at random. Average classification accuracy is reported as a summary statistic over the 100 (10×10) different results. Similarly, the open set experiments make use cross-validation, but with five folds.

4.2 N-Gram Technique

This methodology is used to find the co-occurrence of the words in the sentences of tweets as well as media news and the Outlier detection.

Here is the sentence

1. China is the most populated city in the world . Here key words are **china, most, populated, City, world** No of keywords are 5 let us take it as N. For two gram ,number of loops are N-1

China-most Most-populated populated-City City-world

For three gram ,number of loops are N-2

China-most-populated

populated-City-world

4.3 Cosine Similarity

Step1: Get the two sentences.

Step 2: Extract the words present in the sentences using String split, delimeter is space.

Step 3: Now count the number of times each of those words appears in each sentence.

Step 4: Create the vectors of the count of words of each sentence, lets be the d1 and d2.

Step 5: Multiply the respective place vector of one sentence with another and the do the summation of all, Lets be the d1 * d2.

Step 6: Square the vector 1 values and do the summation, and do the power of 0.5 with that summation, lets be the ||d1||

Step 6: Square the vector 2 values and do the summation, and do the power of 0.5 with that summation, lets be the ||d2||.

Step 6: Calculate cos(d1,d2)=(d1*d2) / (||d1|| * ||d2||)

Example :

Here are two very short texts to compare:

1. Zoha hates me more than Arha hates me
2. Jack likes me more than Zoha hates me

In project want to know how similar these texts are, purely in terms of word counts (and ignoring word order). In project begin by making a list of the words from both texts: each of these word s appears in each text:

Words	Sentence 1	Sentence 2
Me	2	2
Zoha	1	1
Arha	1	0
Jack	0	1
Likes	0	1
hates	2	1
more	1	1
Than	1	1

The information are not interested in the words themselves though the things which are interested in two vertical vector counts

The information can be closed each other by calculating one function of two vectors the cosine angle between them are

D1 : [2,0,1,1,0,2,1,1]

D2 : [2,1,1,0,1,1,1,1]

d1*d2=2*2+0*1+1*1+1*0+0*1+2*1+1*1+1*1=9

||d1||=2*2+0*0+1*1+1*1+0*0+2*2+1*1+1*1=12

||d2||=2*2+1*1+1*1+0*0+1*1+1*1+1*1+1*1=10

Cos(d1,d2)=Cos((d1*d2) / (||d1|| * ||d2||))

Cos(d1,d2)=Cos(9/120)= Cos(0.075);

Cos(0.075)=0.99

Means 99% similar

5. CONCLUSION

The enormous popularity of social media means that it is now a conduit for both legitimate and illegitimate messages targeted at the broadest possible audience. Correspondingly, new forensic challenges have appeared related to this form of new media, triggering the need for effective solutions and requiring the attention of the

information forensics community. A primary problem in this area has been authorship attribution for short messages. In this vein, this study showed that for popular services like Twitter, we face the dilemma of simultaneously having an enormous overall corpus, yet scarcity of information for individual users. This suggests that in project should consider strategies that are a bit different than traditional authorship attribution algorithms for long form writing. When working with highly constrained forms of writing like tweets, the problem size grows rapidly due to the large number of users and messages involved. One way to address this problem is to compute very low-level lexical statistics, which easily leads to high- dimensional spaces. Moreover, the problem is exacerbated by the unconventional punctuation, abbreviations, and character- based signifiers common in Internet culture.

REFERENCES

[1] A. Abbasi and H.Chen. Applying authorship analysis to extremist group web forum messages. IEEE Intelligent Systems, 20(5):67–75, 2005.

[2] 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.

[3] S. Afroz, A. Caliskan- Islam, A. Stolerman, R. Greenstadt, and D. Mc-Coy. Doppelganger finder: Taking stylometry to the underground. In IEEE Security and Privacy, 2014.

[4] J. Albadarneh, B. Talafha, M. Al-Ayyoub, B. Zaqaibeh, M. Al-Smadi, Y. Jararweh, and E. Benkhelifa. Using big data analytics for authorship authentication of arabic tweets. In IEEE/ACM Intl. Conference on Utility and Cloud Computing, 2015.

[5] M. Almishari, D. Kaafar, E. Oguz, and G. Tsudik. Stylometric linkability of tweets. In Workshop on Privacy in the Electronic Society, 2014.

[6] A. Anderson, M. Corney, O. de Vel, and G. Mohay. Multi-topic e- mail authorship attribution forensics. In ACM Conference on Computer Security, 2001.

[7] Yui Arakawa, Akihiro Kameda, Akiko Aizawa, and Takafumi Suzuki. Adding twitter-specific features to stylistic features for classifying tweets by user type and number of retweets. Journal of the Association for Information Science and Technology, 65(7):1416–1423, 2014.

[8] S. Argamon, M. Koppel, and G. Avneri. Style-based text categorization: What newspaper am I reading? In AAAI Workshop on Text Categorization, pages 1–4, 1998.

[9] S. Argamon, C. Whitelaw, P. Chase, S. R. Hota, N. Garg, and S. Levitan. Stylistic text classification using functional lexical features. Journal of the American Society for Information Science and Technology, 58(6):802–822, 2007.

[10] Juola & Associates. Computational analysis of authorship and identity for immigration. Whitepaper, 2016.

[11] H. Azarbonyad, M. Dehghani, M. Marx, and J. Kamps. Time-aware authorship attribution for short text streams. In Intl. ACM SIGIR Conference on Research and Development in Information Retrieval, 2015.

[12] 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.

[13] 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.

[14] D. Bagnall. Author identification using multi-headed recurrent neural networks. In Conference and Labs of the Evaluation Forum, 2015.

[15] A. Bartoli, A. Dagri, An author verification approach based on differential features. CLEF 2015 Evaluation Labs, 2015.

[16] D. Benedetto, E. Caglioti, and V. Loreto. Language trees and zipping. Physical Review Letters, 88(4):048702, 2002.

[17] Y. Bengio. Learning Deep Architectures for AI. Now Publishers, 2009.

[18] M. Bhargava, P. Mehndiratta, and K. Asawa. Stylometric analysis for authorship attribution on Twitter. In Big Data Analytics, 2013.

[19] 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.

[20] J. N. G. Binongo and M. W. A. Smith. The application of principal component analysis to stylometry. Literary and Linguistic Computing, 14(4):445–466, 1999.