Stylometry as a reliable method for fall back Authentication

April 30, 2019

Kishor Datta Gupta
The University of Memphis
Memphis,USA
kgupta1@memphis.edu

Abstract

This Projects aim to evaluate the efficiency of Stylometry approach as an authentication method. Stylometric analysis research has been done for author identification and there is significant progress to recognize an author based on their written texts. In this project, we tried to detect differences between writing styles on the same topic provided by a set of users and we test that these differences are enough to use for an authentication system or not.

1 Introduction

Authentication is a protocol which maintains that only a valid user will be able to get access to a digital asset. Fallback authentication is a support protocol system when users forgot verification credential and need to reset the credentials. Currently, the most popular kind of fallback authentication is some set of security questions. If users able to answer these questions, the system will accept the user as a valid user and reset their credentials.

Security questions are not secure anymore as the current age of social network progress, personal information which used to form reliable questions are now open in public eye, Hackers can easily guess the answer of these questions and can get past the authentication system. We want to apply the stylometric approach to employ a fallback authentication system. Stylometry is the study of writing style. Author Ramaya et el [4] says "There is an unconscious aspect to an author's style that cannot be consciously manipulated but which sesses quantifiable and distinctive features." Based on that we can say with precise feature identification it should be sible to identify the author. But in authentication system, writing sample won't is enough due to usability reason. So it will be a challenge to use it for authentication pure. In this project, we tried to detect feature is small size text sample and later used that for authentication pure. We evaluate our result, and we find that stylometry can be useful with the help of other authentication factors.

2 Related work

In 2014, Sara et al. [1] described four techniques for authorship detections. They were probabilistic models, comparison models, Machine learning classifiers and clustering algorithms and inter-textual distance.

In 2016, T reddy et al.[5] published a paper which summarises the stylometry research for predict the demographic features of authors such as gender, age and personality traits based on the

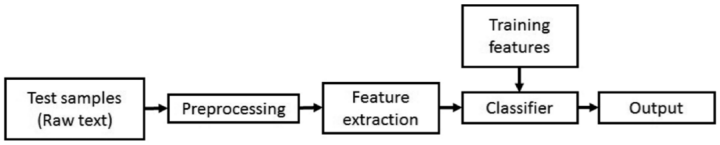


Figure 1: A typical authorship attribution system

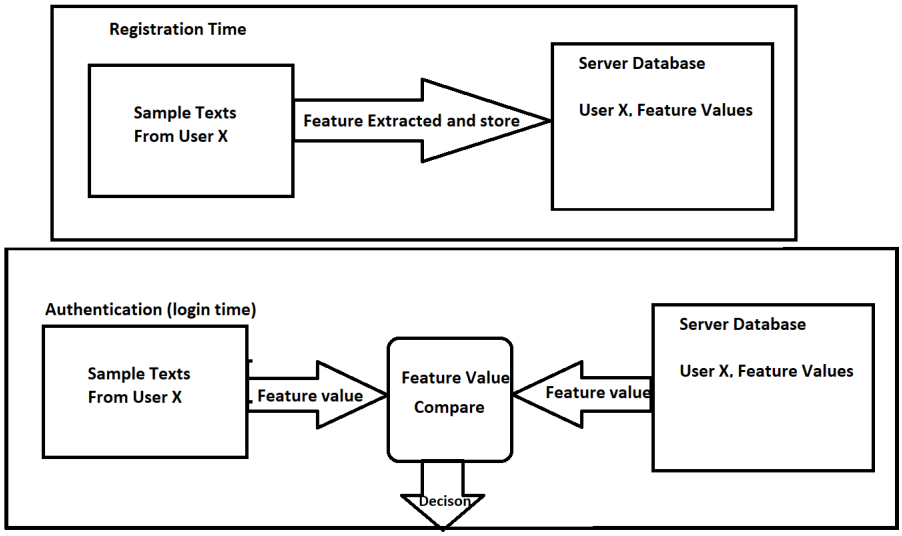


Figure 2: Authentication with stylometry

text corpus written by various authors. In the same year, an analysis of authorship detection using stylometry has done by Neal et al [3]. which studied with 1000 author corpus. In 2017, Mahmoud Khonji et al [2]. published a paper analyzing existing stylometry techniques and introduced the new python library name Fextractor. Which generalizes all existing n-gram based feature extraction methods under the at least l-frequent, 15 dir-directed, k-skipped n-grams, and allows grams to be diversely defined, including definitions that are based on high-level grammatical aspects, such as Part of Speech () tags, as well as lower-level ones, such as the distribution of function words, word shapes, etc.

In fig 1 , A traditional authorship detection system has shown, where some text sample was processed, features are extracted and match with already trained function and output is the percentage of most matched author.

In figure 2, Basic of our authentication system is provided. In Registration time user will provide some text sample for the system to analyze, the system will store analyze reports in server. Later in fallback authentication time, the user will give another sample writing sample. This sample will analyze and compare with saved sample reports if the comparison is above an acceptable threshold system will accept the user as a valid user.

3 Approach

For stylometry, there is lexical, syntactic, semantic, structural, etc. feature. Due to the requirement of authentication approaches we cant have extended text sample and multiple paragraphs. So structural and content specific features are not suitable for our need. We used the 45 features mentioned in figure 3.

As we can not afford to have enough data sample to run a meaning full machine learning classifier due to we are using it for authentication purpose. We have to distinguish between authors based on the variance of feature value.

In our approach, we take the average and standard deviation of all users every feature value.

So if there are

N number of Users are U_1, U_2,U_3.....U_N

and

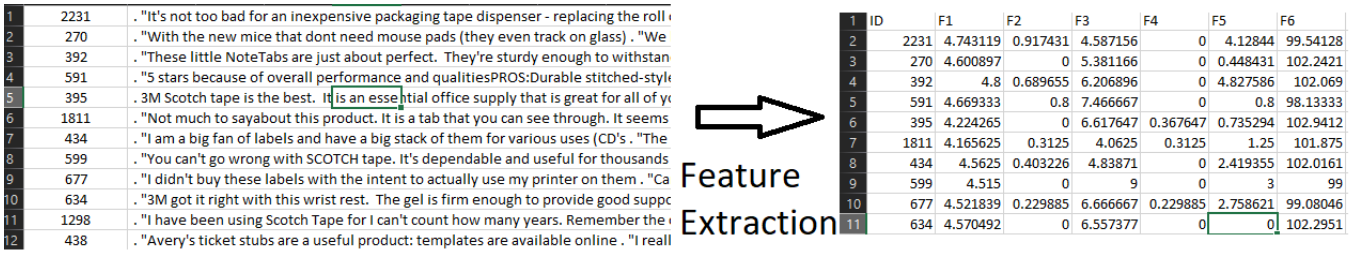
k number of features F_1, F_2,F_3.....F_K

So, each user has feature value as

$$U_1F_1,\ U_1F_2,U_1F_3.....U_1F_K$$
$$U_2F_1,\ U_2F_2,U_2F_3.....U_2F_K$$
$$U_3F_1,\ U_3F_2,U_3F_3.....U_3F_K$$
$$...$$

Feature	Comment
Average length of word	Word length
! Frequency ratio	character feature
? Frequency ratio	character feature
apostephus Frequency ratio	character feature
Semi colon ratio	character feature
Colon ratio	character feature
Cotation marks ratio	character feature
whitespace ratio	character feature
CC pos	Coordinating conjunction
CD pos	Cardinal number
DT pos	Determiner
EX pos	Existential *there*
FW pos	Foreign word
IN pos	Preposition or subordinating conjunctio
JJ pos	Adjective
JJR pos	Adjective, comparative
JJS pos	Adjective, superlative
LS pos	List item marker
MD pos	Modal
NN pos	Noun, singular or mass
NNS pos	Noun, plural
NNP pos	Proper noun, singular
NNPS pos	Proper noun, plural
PDT pos	Predeterminer
POS pos	Possessive ending
PRP pos	Personal pronoun
PRP$ pos	Possessive pronoun
RB pos	Adverb
RBR pos	Adverb, comparative
RBS pos	Adverb, superlative
RP pos	Particle
SYM pos	Symbol
TO pos	*to*
UH pos	Interjection
VB pos	Verb, base form
VBD pos	Verb, past tense
VBG pos	Verb, gerund or present participle
VBN pos	Verb, past participle
VBP pos	Verb, non-3rd person singular present
VBZ pos	Verb, 3rd person singular present
WDT pos	Wh-determiner
WP pos	Wh-pronoun
WP$ pos	Possessive wh-pronoun
WRB pos	Wh-adverb

Figure 3: Feature list



1	2231	. "It's not too bad for an inexpensive packaging tape dispenser - replacing the roll
2	270	. "With the new mice that dont need mouse pads (they even track on glass) . "We
3	392	. "These little NoteTabs are just about perfect. They're sturdy enough to withstan
4	591	. "5 stars because of overall performance and qualitiesPROS:Durable stitched-styl
5	395	. 3M Scotch tape is the best. It is an essential office supply that is great for all of y
6	1811	. "Not much to sayabout this product. It is a tab that you can see through. It seems
7	434	. "I am a big fan of labels and have a big stack of them for various uses (CD's . "The
8	599	. "You can't go wrong with SCOTCH tape. It's dependable and useful for thousands
9	677	. "I didn't buy these labels with the intent to actually use my printer on them . "Ca
10	634	. "3M got it right with this wrist rest. The gel is firm enough to provide good suppo
11	1298	. "I have been using Scotch Tape for I can't count how many years. Remember the
12	438	. "Avery's ticket stubs are a useful product: templates are available online . "I reall

	ID	F1	F2	F3	F4	F5	F6
2	2231	4.743119	0.917431	4.587156	0	4.12844	99.54128
3	270	4.600897	0	5.381166	0	0.448431	102.2421
4	392	4.8	0.689655	6.206896	0	4.827586	102.069
5	591	4.669333	0.8	7.466667	0	0.8	98.13333
6	395	4.224265	0	6.617647	0.367647	0.735294	102.9412
7	1811	4.165625	0.3125	4.0625	0.3125	1.25	101.875
8	434	4.5625	0.403226	4.83871	0	2.419355	102.0161
9	599	4.515	0	9	0	3	99
10	677	4.521839	0.229885	6.666667	0.229885	2.758621	99.08046
11	634	4.570492	0	6.557377	0	0	102.2951

Figure 4: Feature Data

Range	F1	F2	F3	F4	F5	F6
AVG	4.583513	0.475856	5.427911	0.107806	1.956396	101.3563
STD	0.295095	0.557469	2.22014	0.227633	1.389775	2.127606
UR	5.173703	1.590794	9.868191	0.563073	4.735947	105.6115
LR	3.993323	-0.63908	0.987632	-0.34746	-0.82315	97.10107

Figure 5: Bound Data

$$\dots$$
$$U_N F_1,\ U_N F_2, U_N F_3 \dots\dots U_N F_K$$

Now for every feature value f_J we will get

$$f_{j_{avg}} = \frac{\sum_{i=0}^{N} U_i F_j}{N}$$

Now every feature value f_J standard deviation is σf_j
We will determine upper bound and lower bound as below

$$\text{Upper bound } U_b = f_{j_{avg}} + \sigma f_j$$
$$\text{Lower bound } L_b = f_{j_{avg}} - \sigma f_j$$

Now for every user we will create profile P_i
P_i has sequence of feature value in $+1, -1, 0$.
if a User U_i has feature value F_j, profile value of user for that feature will be as below

$$\text{if } F_j > U_b\ , P_i = 1$$
$$\text{if } F_j < L_b\ , P_i = -1$$
$$\text{if } U_b > F_j > L_b\ , P_i = 0$$

In registration time we will store $P_{i_{REG}}$ values for each users provided text sample, and in login time we will again generate the $P_{i_{login}}$ from provided sample and try to match both.
if for more than 50% of total feature F_K

$$P_{i_{REG}} \equiv P_{i_{login}}$$

we will consider it is a matched user.

4 Experiments & Results

For my experiment, I used the Amazon product review data set, where the same users provided a review on similar products. We randomly took 63 reviewers, all of them have more than 7 product review. And all products are office suppliers tools. This way we can make each text domain specific. No analysis is more than two lines.

I first remove 1 of the review as a test sample from each of the users and start our process on remaining reviews. We extracted 45 feature. Thirty-six of them are the ratio of POS tag usage. Others are some lexical features. I used c-sharp and a wrapper of Stanford NLP library to generate feature data. We take 45 feature as shown in figure 3. After I made the feature data using the math function, I calculate the average and standard deviations of feature value. With that, I created a profile for each user using c-sharp services and saved them in a CSV file. During login time, I generated profile value in the same way and checked the match percentage with the userid associated with the sample text. Based on that percentage login successful and failure has decided.

	ID	P1	P2	P3	P4	P5	P6
2	2231	0	0	1	0	0	0
3	270	0	0	0	0	0	0
4	392	0	1	0	0	1	0
5	591	1	0	-1	0	0	0

Figure 6: Profile Data

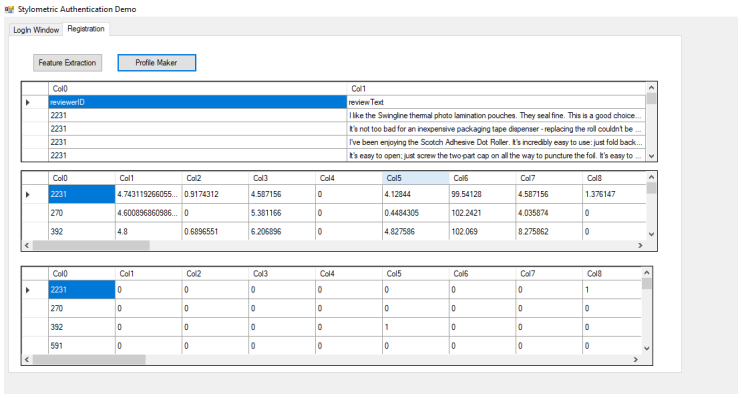


Figure 7: Demo Application

In figure 4,the extracted feature data has showed. From all of these feature data, i generated upper bound and lower bound of feature which will use for as showed in figure 5. Based on the bound data, we generated profile information of each users as we described in our approach. Sample profile data has showed in figure 6. Above way I developed a database with userid and their profiles id, For testing purpose, I developed a c-sharp winform application as showed in figure 7.

In our experiment, I found a significant percentage of time writing style can be matched with profile values above 50%. But with cross-testing, I saw it don't have the performance accuracy I need for an authentication purpose.

But the users who have larger text sample has more accuracy. And the deviation of profile values. So it may be possible that if we add more features. This approach can be used in an authentication system.

5 Conclusions and future work

From my experiment, it is evident that the current approach has not enough unique features to correctly identify authors. But combining with more strategy and adding new features it may be possible to increase the efficiency. In the current state, it is not reliable method to use in authentication system. In my future work, I want to add other approach combining with the correct and try to improve the system.

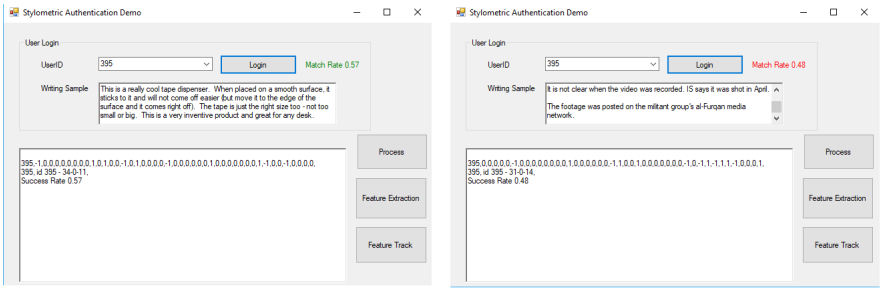


Figure 8: Login Testing

References

[1] S. E. M. El and I. Kassou. Authorship analysis studies: A survey. *International Journal of Computer Applications*, 86(12), 2014.

[2] M. Khonji and Y. Iraqi. De-anonymizing authors of electronic texts: A survey on electronic text stylometry. 2017.

[3] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, and D. Woodard. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSUR)*, 50(6):86, 2018.

[4] C. H. Ramyaa and K. Rasheed. Using machine learning techniques for stylometry.

[5] T. R. Reddy, B. V. Vardhan, and P. V. Reddy. A survey on authorship profiling techniques. *International Journal of Applied Engineering Research*, 11(5):3092–3102, 2016.