Journal of Forensic Sciences And Criminal Investigation
ISSN: 2476-1311

Juniper Publishers key to the Researchers

Research Article
Volume - 8 Issue - April 2018
DOI: 10.19080/JFSCI.2018.08.555733

J Forensic Sci & Criminal Inves

Authorship Profiling of Instant Messaging Sites based on Stylistic and Stylometric Analysis

Gloria Christal[1], Prajakta Manve[1], Pooja Ahuja*[2] and MS Dahiya[3]

1Student, MSc. Forensic science, Institute of Forensic Science, Gujarat Forensic Sciences University, India

2Assistant professor, Institute of Forensic Science, Gujarat Forensic Sciences University, India

3Director, Institute of Forensic Science, Gujarat Forensic Sciences University, India

Submission: March 16, 2018; **Published:** April 04, 2018

***Corresponding author:** Pooja Ahuja, Assistant professor, Institute of Forensic Science, Gujarat Forensic Sciences University, India, Email: pahuja159@gmail.com

Abbreviations :IM: Instant Messaging; WST: Wordsmiths Tools

Introduction

The increase in popularity of the Internet media, like emails, blogs/internet forum and websites have been identified as the ideal communication platform for people and one such medium is Instant Messaging (IM) which has gained prominence recently with rise of the Internet. Instant messaging is a type of online chat that offers real-time text transmission over the Internet. The Global Web Index report was conducted across 32 markets involving 170,000 internet users. The study shows that 52 per cent of Indian instant messaging users are on WhatsApp, while 42 per cent use Facebook Messenger, 37 per cent use Skype, We Chat has a 26 per cent share in the market and Viber with 18 per cent market share is in the fifth spot IM is a set of communication technologies used for text-based communication between two or more participants over the Internet or other types of networks instantly [1].

Forensic linguistics, legal linguistics, or language and the law, is the application of linguistic knowledge, methods and insights to the forensic context of law, language, crime investigation, trial, and judicial procedure. Applications of forensic linguistics include voice identification, interpretation of expressed meaning in laws and legal writings, analysis of discourse in legal settings, interpretation of intended meaning in oral and written statements, authorship identification and interpretation and translation when more than one language must be used in a legal context." Forensic stylistics is the application of the science of linguistic stylistics to forensic contexts. Common features of style include the use of dialogue, including regional tones and pronunciation and individual dialects (or ideolects), the use of grammar which includes the observation of active voice and passive voice, the use of particular language registers, the distribution of sentence lengths and etc. Stylometrics is a development of literary stylistics, which is based on the assumption that all authors have individual writing habits. These writing habits can be seen in features such as core vocabulary use, phraseology and sentence complexity and all these features are unconscious habits which are well ingrained. Furthermore, it is also concerned with locating textual features which can be used for determining authorship of a text/ writings. This is achieved by having a sample of known authored texts from different authors which can be compared to a disputed text. Stylometrics is a quantitative analysis [2-5].

Thus, Stylometric approaches seek to find or describe quantifiable markers of authorship, which in the general sense vary more between authors than within authors. Typical stylometric markers include relative frequencies of different word classes or even non word letter clusters. The style markers can be categorized as character-based, word-based, sentence-based, document based, structural or syntactic. A few examples of style markers include: function word usage (common adverbs, auxiliary verbs, conjunctions, prepositions and pronouns); word collocations; sentence length and punctuation. Author identification, is the task of determining the author of a piece of work. There are following two types of variation that a forensic expert faces Intra-author variation refers to the ways in which an author's text differs from another text written by the same author, whereas inter-author variation refers to the ways texts vary between different authors. However, in this study, the only causes of variation that could have any bearing are time lapse, if some time has passed between posts, and change in circumstances if the writer has undergone any recent changes in her life. Despite Facebook being a social networking site, there could be socio metric parameters, as it is common for an individual to have on

their friends' list people who occupy different power positions, and that will affect the choice of language. The challenges for applying the authorship identification to online messages are

a) Length of Online document is short.

b) An online document has less formal writing style and the vocabulary pattern is not stable.

c) Online documents are different than normal text documents in composition style and in format of structure.

d) Due to the internationalization of cybercrime, multilingual problems become a new challenge for authorship Analysis. In an authorship attribution context, quantification refers to the identification and frequency counting of selected linguistic features (style markers), which are then statistically measured in order to determine the origin of a text. Moreover, statistically measuring frequency counts would provide the researcher with the evidence of whether the differences are significant, i.e. whether there is less than a 5% or 1% chance of a specific feature or group of features occurring randomly [6-8].

Method and Methodology

Twenty-five subjects for this study were matched as closely as possible according to factors such as age, race, class and education level. Both male and female subjects, aged between 20 and 30, who were educated till undergraduate level in India, were taken. They spoke English as a second language and Hindi/ regional language as first language and were active users of social networking and instant messaging sites such as WhatsApp for quite some time. The participants names were renamed, as A to Z, with the disputed text being referred to as Y to maintain the confidentiality of the information. In order to obtain the texts, a message was sent to all the potential candidates on messenger and WhatsApp. A request was made to them asking for 1,000 words of text from their messenger and WhatsApp inbox, cut and pasted onto a word document and e-mailed to me. The text had to be their own typing with no third-party submissions and no editing of the text before submitting it. They were asked to start from their latest text and move backwards until 1,000 words had been obtained so as to avoid any significant time lags and to have writings which were as current as possible and this also helped ensure that all submissions were from the same period of time. One of the participants was asked to submit an extra 1,000 words which would act as the disputed text Y. the contents of the texts weren't altered, apart from removing the names, contact numbers and electronic addresses.

Step 1: The IM conversations were logged to text files and doc or docx format in the following format:

[Timestamp] [User name:] [Message]

The data was subjected to a series of pre-processing steps. The data were prepared for analysis by removing both the timestamp and username. Thus, an example of a formatted log for User L.

Step 2: Next step was to assemble all questioned and known writings and check for compatibility and comparability. Features were chosen using criteria:

a) Deviations from any norm such as errors or mistakes; and

b) Variation within the writer's norm (i.e. does the writer use more than one form in a text (h/hai/hen)).

Categorization of Stylistic Features

The stylistic analysis of each participant's writings is divided into two parts. Noticeable features are extracted and put into tabular form. The table has two distinct parts. The first shows the stylistic features shared with the writer of the disputed text (Y). Therefore, the shared features are not the same for all the writers. The second parts of the table highlight stylistic features which are not shared with the writer of the disputed text Y (Table 1).

Table 1: Stylistics characteristics of writer A and writer Y.

Features	Examples	Frequency	
Features shared by writer y and writer a		A	Y
Typographic misspellings	Pt-pr,(bhau,bhasi,bahi-bhai),jeetu- jeeta,mbe-mne, Grammatical-totally	7	14
Shortenings	Jan, feb, dec, max, min, msgs, bro, lang, dr, ira icp,np,xp,pc,nrml,gui,bipm,sdlc	18	21
Typographic characterstics	•Inappropriate punctuation Double questionmark(there??)	1	2
emoticons	,	2	2
Exclamatory	ah, oh shit!!	2	2
Repetition of words	Ha ha, sry sry,acha acha,okay okay	4	2
Homophone substitution	M,r,u,b,n	5	3
One word substituted with multiple spellings	Ok(okay,ohkay,ohkayz,ohkda),bro bhai	2	2

How to cite this article: Gloria C, Prajakta M, Pooja A, MS Dahiya. Authorship Profiling of Instant Messaging Sites based on Stylistic and Stylometric Analysis. J Forensic Sci & Criminal Invest. 2018; 8(2): 555733. DOI: 10.19080/JFSCI.2018.08.555733

Characters not shared by writer a and y			
Inappropriate capitalization	Na,apn,jeetu,jo,teri	5	0
Initial capitalization	Monday,dec,java,jan,feb,I, google,	7	0
Digitally	Hahaha,yo(yes),bubye	3	0
mediated			
communications			
Contains farewell	Goodnight, Bubye	2	0
Use of other language	Punjabi –vaari,liyo,utta,kariyo	5	17
Does not contain prosodic emphasis			
Typographic characterstics	•Word+ellipses(…)+word(is…formatting		
Emoticons	•Apostrophic coma present(it's)		
	•Inappropriate punctuation		
	1-Word+fullstop+word(thode.se)		
	2-double exclamatory mark(oh shit!!		
	•All English words are spelled correctly		

Step 3: The quantitative analysis, where programs like Wordsmiths Tools (WST) and text stat tool and Microsoft word to count the frequencies of function words, frequently occurring words and punctuation marks and to look for keywords, followed by a Chi-square test of the profiles comparing the questioned text to each of the known texts (Table 2).

Table 2: General characteristics of writer Y.

Features	Y
Characters	4665
Lines	299
Words	1000
paragraphs	299
Numerals	17
punctuations	2
Average word length(char)	3.5
Average sentence length (word)	3.5
No of words per sentence	3.47
Unique words	435(43%)
Monosyllabic words	560
Polysyllabic words(>3)	73
Syllables per word (approximate)	1.5
Short words (<3)	510(51%)
Long words(>7)	44(4%)
Longest sentence (by no of characters)	61 char,31 words
Longest word(by no of characters)	10 char,3 words
Lexical density	43.41
Flesch index	80.55

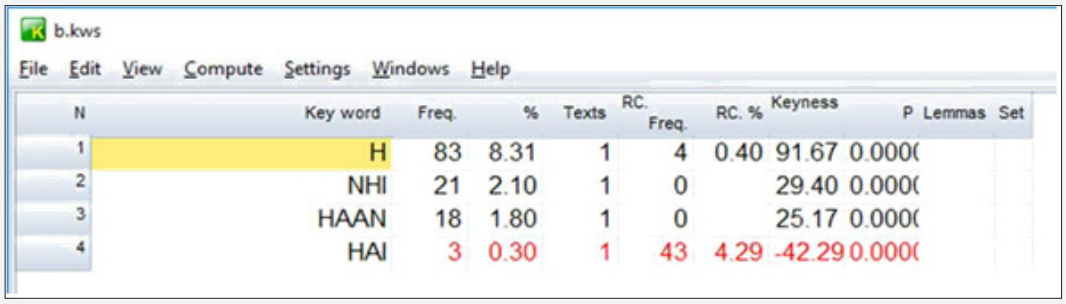
Table 3: Extract from the table used to analyse function words.

Words	y	a	B
0r/aur	18	9	6
Se	12	9	9
Bhi	9	21	12
Gaya/gya	13	6	6
Ka	13	4	16
Kya	21	10	7
Ke liye	3	0	3
Tu/tum	8	7	3
Me/mai/mein	13	20	19
Yeh	0	7	0
Hai/y	43	42	86
Toh/to	12	48	28
Pr/per/pe	11	7	2
Ki	10	6	10
Is	2	1	4
Na	6	10	3
Koi	5	8	5
Us	1	0	0
Ya	2	4	0
sab	1	0	5
The	1	0	6
tha	7	8	8
Ha/haan	8	30	19
Nahi/nai/nhi	13	8	22

How to cite this article: Gloria C, Prajakta M, Pooja A, MS Dahiya. Authorship Profiling of Instant Messaging Sites based on Stylistic and Stylometric Analysis. J Forensic Sci & Criminal Invest. 2018; 8(2): 555733. DOI: 10.19080/JFSCI.2018.08.555733

Test 2: Keywords: "Key-words provide a useful way to characterize a text or a genre" (Figure 1).

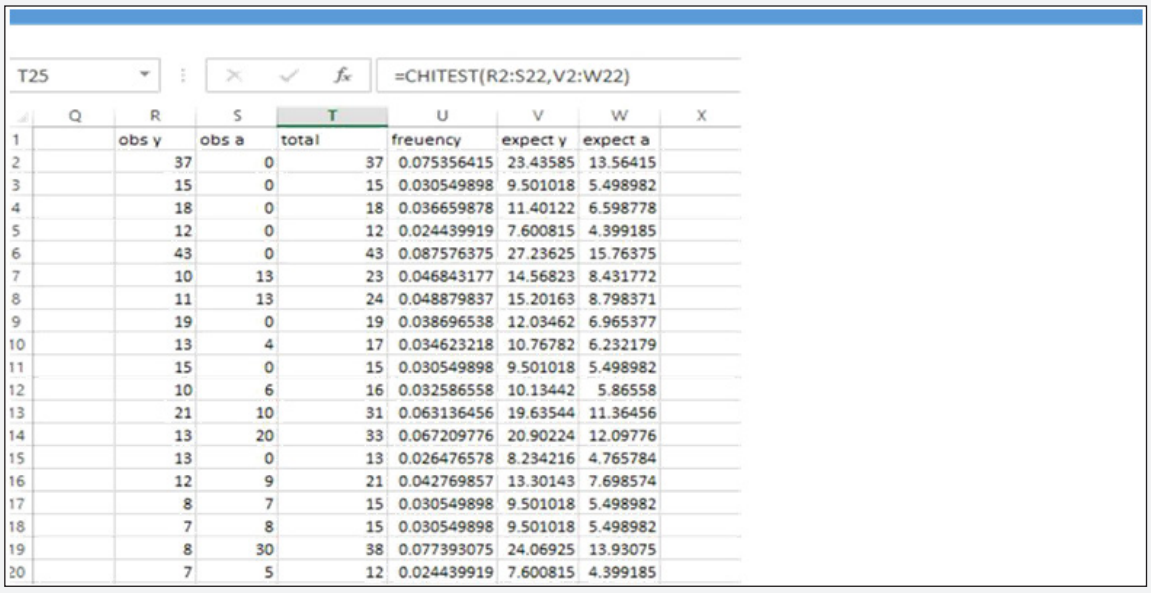
Test 3: Function Words: Wordlist has a number of features useful to a forensic linguistic study, Firstly; it generates word listing in alphabetical order and/or frequency order, so texts can be analyzed at a lexical level. The test involves an analysis of 24 function words (Table 3).



N	Key word	Freq.	%	Texts	RC. Freq.	RC. %	Keyness	P	Lemmas	Set
1	H	83	8.31	1	4	0.40	91.67	0.000(		
2	NHI	21	2.10	1	0		29.40	0.000(		
3	HAAN	18	1.80	1	0		25.17	0.000(		
4	HAI	3	0.30	1	43	4.29	-42.29	0.000(		

Figure 1: Example of keyword test.

The statistical program Excel is used for the Chi-square test calculations needed for the function words, most frequently occurring words and punctuation. Table 2 shows an example of the Chi-square test calculations conducted by excel with regard to the disputed text (Y) and Writer A. function words which were omitted from the calculations, as their frequency counts did not add up to 10. The 1,000-word corpora were subjected to a Chi-square test using the statistics program Excel. The profile of disputed text (Y) is compared to each of the other writers' profiles for these words and the overall results are placed at the bottom of the table below the corresponding author (Figure 2).



	Q	R	S	T	U	V	W	X
1		obs y	obs a	total	freuency	expect y	expect a	
2		37	0	37	0.075356415	23.43585	13.56415	
3		15	0	15	0.030549898	9.501018	5.498982	
4		18	0	18	0.036659878	11.40122	6.598778	
5		12	0	12	0.024439919	7.600815	4.399185	
6		43	0	43	0.087576375	27.23625	15.76375	
7		10	13	23	0.046843177	14.56823	8.431772	
8		11	13	24	0.048879837	15.20163	8.798371	
9		19	0	19	0.038696538	12.03462	6.965377	
10		13	4	17	0.034623218	10.76782	6.232179	
11		15	0	15	0.030549898	9.501018	5.498982	
12		10	6	16	0.032586558	10.13442	5.86558	
13		21	10	31	0.063136456	19.63544	11.36456	
14		13	20	33	0.067209776	20.90224	12.09776	
15		13	0	13	0.026476578	8.234216	4.765784	
16		12	9	21	0.042769857	13.30143	7.698574	
17		8	7	15	0.030549898	9.501018	5.498982	
18		7	8	15	0.030549898	9.501018	5.498982	
19		8	30	38	0.077393075	24.06925	13.93075	
20		7	5	12	0.024439919	7.600815	4.399185	

Figure 2 : Screenshot of Chi Square Calculated In Excel Sheet.

Test 4: The fourth test is an analysis of the most frequently occurring words and is, in certain respects, an extension of function words and work with keywords. The test identified the 24 most frequently occurring words in the disputed text (X), and compared them to the frequencies of same words in the other five texts and put them into tabular form, (Table 4).

Test 5: Punctuation- The fifth test involves analyzing punctuation. As noticed in the stylistic assessment, the writers in this study exhibit numerous non-standard punctuation styles, with multiple question marks, exclamation marks, ellipsis and question mark/exclamation mark, and in addition to that there were several emoticons(:-), :-P, etc.). Therefore, I categorized the features as sets, for example double question marks and triple question marks were taken separately. For this analysis I used the search function of Microsoft Word. This function has the advantage of being able to display all of the features from the original document as they are highlighted in the text and then counted manually (Table 5).

How to cite this article: Gloria C, Prajakta M, Pooja A, MS Dahiya. Authorship Profiling of Instant Messaging Sites based on Stylistic and Stylometric Analysis. J Forensic Sci & Criminal Invest. 2018; 8(2): 555733. DOI: 10.19080/JFSCI.2018.08.555733

Table 4: Extract from the most frequently occurring words.

SL.NO	WORD	Y	a
1.	acha	37	0
2.	are	15	0
3.	aur	18	0
4.	bata	12	0
5.	Hai	43	0
6.	Ho	10	13
7.	Hi	11	13
8.	k	19	0
9.	ka	13	4
10.	kar	15	0
11.	ki	10	6
12.	kya	21	10
13.	me	13	20
14.	nahi	13	0
15.	se	12	9
16.	ab	7	0
17.	Tu	8	7
18.	Tha	7	8
19.	Ha	8	30
20.	Gaya	9	0
21.	Bhi	9	0
22.	bhai	7	5
23.	toh	12	47
24.	fir	7	8

Table 5: Extract of the table used to analyse punctuation.

Punctuations	Writer Y	Writer A	Writer B	Writer C
Full stop	0	7	2	1
Comma	0	0	1	0
Apostrophic mark	0	1	1	0
Question mark				
One(?)	0	0	20	0
Two(??)	1	2	0	1
Three(???)	2	0	0	2
Ellipses				
2 dots(..)	0	0	1	0
3 dots(...)	0	2	3	2
4 dots(....)	0	0	26	1
asterisks	0	0	1	1
Exclamation marks				
1	0	0	0	0
2	0	0	0	0
3	0	1	0	0
Ampersand	0	0	0	0
hyphen	2	0	1	1
Colon	0	0	0	1
emoticons				
	0	6	0	0
	2	2	4	4
	2	0	7	4
p-value		0.10	0.00	0.81

Results and Discussions

Stylistic Analysis

Table 6 All the candidate texts (Writers A-Z) were analyzed stylistically and individually against the disputed text Y, with the purpose of finding the features shared with Y and the features not shared with Y. All the candidate authors shared features with Y, as well as exhibiting various features that were not shared with Y. Only Writer C exhibited meaningful similarities and very few dissimilarities, which led to the conclusion that it was highly probable that C is the writer of the disputed text Y.

How to cite this article: Gloria C, Prajakta M, Pooja A, MS Dahiya. Authorship Profiling of Instant Messaging Sites based on Stylistic and Stylometric Analysis. J Forensic Sci & Criminal Invest. 2018; 8(2): 555733. DOI: 10.19080/JFSCI.2018.08.555733

Table 6: Result of Stylistic Analysis of Writer A-Z Compared To Y.

SUSPECT/AUTHOR	Features shared	Features unshared
Writer Y/A	8	8
Writer Y/B	9	8
Writer Y/C	13	1
Writer Y/D	10	8
Writer Y/E	8	6
Writer Y/F	9	9
Writer Y/G	9	9
Writer Y/H	10	10
Writer Y/I	9	10
Writer Y/J	10	10
Writer Y/K	8	7
Writer Y/L	9	8
Writer Y/M	10	9
Writer Y/N	9	8
Writer Y/O	9	7
Writer Y/P	9	8
Writer Y/Q	9	11
Writer Y/R	10	9
Writer Y/S	9	8
Writer Y/T	8	12
Writer Y/U	9	9
Writer Y/V	10	8
Writer Y/W	7	8
Writer Y/X	9	7
Writer Y/Z	7	10

Keywords

Table 7 This section examines the keywords at the between the reference corpus (Y) and each of the study corpora in turn (A-Z). The above table clearly shows that Writer C is correctly clustered with Writer Y, having the lowest keywords and lowest average keyness. Keyword analysis utilised a chi square test to determine the keyness value of identified lexical items in the two texts in order to prove the hypothesis or otherwise. If the average keyness value (in particular of grammatical words) were to be lower than 10, it could be concluded with a reasonable measure of confidence that there was a strong likelihood that the text was produced by the same author (or joint authors).

Table 7: Result of Keyword Analysis of Writer A-Z Compared To Y.

Suspect/Author	Key words	Average keyness value
Y/A	4	39.96
Y/B	4	47.13
Y/C	0	0
Y/D	4	29.45
Y/E	3	41.56
Y/F	5	45.62
Y/G	1	27.87
Y/H	2	39.04
Y/I	3	35.36
Y/J	3	40.23
Y/K	1	38.19
Y/L	5	36.38
Y/M	4	37.21
Y/N	2	39.22
Y/O	2	51.24
Y/P	2	49.915
Y/Q	2	30.68
Y/R	2	54.97
Y/S	2	40.045
Y/T	2	38.58
Y/U	3	44.836
Y/V	3	47.46
Y/W	3	43.143
Y/X	4	38.805
Y/Z	4	52.317

General Characteristics

(Table 8). This section examines the general characteristics of all the writers (A-Z). The above table clearly shows that Writer C is and Writer Y have almost same average value thus this result can be used for the narrowing down of the list of suspects. This test should be treated with a fair degree of caution as it has many red flags as the values of writer A, Writer D are also very close to writer Y.

How to cite this article: Gloria C, Prajakta M, Pooja A, MS Dahiya. Authorship Profiling of Instant Messaging Sites based on Stylistic and Stylometric Analysis. J Forensic Sci & Criminal Invest. 2018; 8(2): 555733. DOI: 10.19080/JFSCI.2018.08.555733

Table 8: Result of General Characterstics Analysis of Writer A-Zcompared to Y.

Suspect/Author	Average value
A	**428.706**
B	443.523
C	428.563
D	423.117
E	430.095
F	440.844
G	449.962
H	466.16
I	466.925
J	484.698
K	460.723
L	416.069
M	484.577
N	506.603
O	418.003
P	469.348
Q	458.880
R	446.340
S	451.219
T	443.07
U	415.011
V	443.28
W	429.12
X	509.251
Y	423.891
Z	416.402

Table 9: Result of Function Words Analysis of Writer A-Z Compared To Y.

Suspect/Author	p-value
A	0.00
B	0.522
C	0.0003
D	0.029
E	0.00
F	0.00
G	0.0795
H	0.0005
I	0.0002
J	0.00
K	0.00
L	0.00
M	0.00
N	0.0600
O	0.00
P	0.00
Q	0.00
R	0.0003
S	0.0033
T	0.0159
U	0.0001
V	0.00
W	0.0005
X	0.0003
Z	0.00

Function Words

Table 9 Ho-there is no difference between usage of function words of the writers i.e. hypothesis of independent

H1: there is difference between usage of function words of the writers i.e. hypothesis of dependent or hypothesis of sameness.

Ho: null hypothesis is (accepted if p value<0.05).

H1: alternate hypothesis (accepted if p value >0.05).

It can be seen that writers C, G and N meet the requirements of the hypothesis of sameness as they all exhibit p>0.05 (Chaski 2001). However, a closer analysis reveals some considerable differences. In order of furthest to closest to Y, it can be seen that N is the furthest away from Y, followed by G and C. C is the closest to X with a p value of 0.522.

Frequently Occurring Words

Table 10 Ho: there is no difference between usage of frequently occurring words of the different writers i.e. hypothesis of independent.

H1: there is difference between usage of frequently occurring words of the different writers i.e. hypothesis of dependent or hypothesis of sameness Ho null hypothesis is (accepted if p value<0.05)

H1: alternate hypothesis (accepted if p value >0.05).

Out of 25 writers only writer C met the criterion for the hypothesis of sameness with a p value of 0.06. Thus, this style marker was quite effective at identifying the author of the 'disputed text 'unlike function words where more writers are fulfilling the criterion for the acceptance of hypothesis of sameness.

How to cite this article: Gloria C, Prajakta M, Pooja A, MS Dahiya. Authorship Profiling of Instant Messaging Sites based on Stylistic and Stylometric Analysis. J Forensic Sci & Criminal Invest. 2018; 8(2): 555733. DOI: 10.19080/JFSCI.2018.08.555733

Table 10: Result of Frequently Occuring Words Analysis of Writera-Z Compared To Y.

Suspect/Author	p-value
A	0.00
B	0.00
C	0.06
D	0.00
E	0.00
F	0.00
G	0.00
H	0.00
I	0.00
J	0.00
K	0.00
L	0.00
M	0.00
N	0.00
O	0.00
P	0.00
Q	0.00
R	0.00
S	0.00
T	0.00
U	0.00
V	0.00
W	0.00
X	0.00
Z	0.00

Punctuations

Table 11 Ho: there is no difference between usage of punctuation analysis of the different writers i.e. hypothesis of independent

H1: there is difference between usage of punctuation analysis of the different writers i.e. hypothesis of dependence or hypothesis of sameness Ho null hypothesis is (accepted if p value<0.05).

H1: alternate hypothesis (accepted if p value >0.05).

Table 11: Result of Punctuation Analysis of Writer A-Z Compared to Y.

Suspect/Author	p-value
A	0.017
B	0.00
C	0.87
D	0.00
E	0.00
F	0.018
G	0.130
H	0.00
I	0.00
J	0.00
K	0.00
L	0.00
M	0.00
N	0.00
O	0.0002
P	0.00
Q	0.080
R	0.00
S	0.00
T	0.002
U	0.00
V	0.00
W	0.001
X	0.00
Z	0.086

The most noticeable aspect of this table is the different types of punctuation employed by the 25 authors, from the standard prescriptive use of punctuation to the more creative examples, such as using multiple question marks and emoticons. Many of the features were only used by one or two authors, which resulted in many frequency counts being too low to be statistically viable. Turning now to the significance levels between the profiles of writers A-Z compared to Y, it can be seen that C met the p>0.05 criterion (Chaski 2005) with a p value of 0.87. The profiles of the others were very significantly different to Y [9].

Conclusion

This study shows that highly literate writer is also able to use informal language based on spoken registers when they feel like it, following a trend in digitally mediated communication

How to cite this article: Gloria C, Prajakta M, Pooja A, MS Dahiya. Authorship Profiling of Instant Messaging Sites based on Stylistic and Stylometric Analysis. J Forensic Sci & Criminal Invest. 2018; 8(2): 555733. DOI: 10.19080/JFSCI.2018.08.555733

for social purposes in general. The first objective of the study was to explore the texts stylistically by analyzing patterns in punctuation, typography, spelling, language used and features associated with digitally mediated communication in the disputed texts. It can be concluded that punctuation is a style marker that correctly identified Writer C as the most probable author of the disputed text. A possible reason for success of punctuation style marker is nature of punctuation usage in digitally mediated communication, which is characterized by far greater freedom for idiosyncratic use, and the fact that there is more likelihood of having more countable features but it should be treated with degree of caution as result would be different writing has used medium less tolerant. Also it is possible to attribute authorship to considerable extent on the basis of use of home language and education level e.g. Most of the participant belonging to Gujarat region have the habit of writing '' che'' as ''6'' which is not very prevalent in other non-Guajarati participant attempting to write in Gujarati language. There is considerable evidence that the methods employed in this study are effective in identifying the writer of disputed text correctly [10-12].

It identifies that creativity in texting is not limited to variation in spelling, but can be seen also in creative patterns reflecting spoken everyday creativity which seems to play a significant role in allowing users to express themselves despite the restrictions of the medium. In the absence of other resources, such as prosody and tone, and without space for the lengthy expression possible in extended written prose, texters exploit and transfer spoken creativity across to texting in order to fulfil expressive, speech-like functions. Language used in messages reveals the presence of spoken features in text sample, it becomes evident when the samples were analysed closely for usage of certain characteristics like punctuations or digitally mediated communication etc. For example, some ellipsis reflects that occurring in spoken language, texters also produce utterances, while many text messages are speech-like in their clausal combination, others combine clauses into short, disjointed sentences; and while there is deixis in texting, it largely relates to time rather than place. Writer C is and Writer Y have almost same average value thus this result can be used for the narrowing down of the list of suspects. This test should be treated with a fair degree of caution as it has many red flags as the values of writer A, Writer D are also very close to writer Y. Thus none of the individual attributes identified by attribute selection were strong enough to determine author identification with a high degree of accuracy.

The keyword test correctly identified writer C as the most likely author of the disputed text as he had no keywords and average keyness value zero in 100 frequently occurring words, however writer G and writer K have one keyword which put them close to writer C due to this keywords need to be used with a caution when analysing short text. Function words are not very successful of the style markers as 5 out of 25 writers met the criterion p>0.05. Despite of that, the function word test did identify writer C as the most probable writer of the disputed text with p value 0.522 but with writer G and N with p value 0.079 and 0.06 respectively, so there is doubt as to how effective this style marker is at 1,000 words. The most frequently occurring word is most promising out of all style markers as writer C is correctly identified as most probable writer of the disputed text with p value 0.06. It is better prospect as words come directly from the participants and there by remove the need to exclude certain words due to genre influence. The punctuation style marker test identify writer C as the most probable author of disputed text with p value of 0.86 but writer G, Q and Z with p value 0.13, 0.08 and 0.08 respectively also fulfil the criterion of p value> 0.05 so there is doubt on the effectiveness of this style marker in stylometric analysis, a possible reason for this doubt is that the writers have not used a well versed standard of punctuation and they chose nonstandard forms of punctuation on instant messaging.

Thus the stylometric analysis of the three style markers i.e. Function word, frequently occurring words, punctuations all correctly identified writer C as the author of the disputed text. The order of effectiveness with reference to p values is as follows:

I. Frequently occurring words,

II. punctuations,

III. function words.

Thus, it can be concluded that the most frequently occurring words in a corpus provide a "fingerprint i.e. 'an identifiable unique or distinctive way in which each individual speaks and writes'" for the instant messaging. With the advent of instant messaging and whats app in particular, a whole new world of possibilities has opened up for forensic linguistic research. This is especially so if one considers that whats app is nearing one billion subscribers and has become a political force for social change, with politicians using it for campaigning, Nearly 300 WhatsApp groups were being used to mobilise stone-pelters in Kashmir to disrupt security forces and it to spread its message and disaffected youths in Kashmir. It is my hope that this study has made some contribution to the field of authorship attribution in general and how it applies to instant messaging such as whats app in particular. The success achieved in this study in exploring the effectiveness of author attribution derives partially from the fact that both stylistic and stylometric characteristic analyses were used.

Limitations of The Study

One of the limitations of the study is that if the time lapse between the samples is more the above criterion for identification can change. Similarly, it is known that variation exists between dialects in terms of both grammatical structures and lexical items. With this in mind, is it can be expected that a person who moves from one geographical location will add dialectal traits from the later location to their current active vocabulary thus varying the way in which they type a text message [12-15].

How to cite this article: Gloria C, Prajakta M, Pooja A, MS Dahiya. Authorship Profiling of Instant Messaging Sites based on Stylistic and Stylometric Analysis. J Forensic Sci & Criminal Invest. 2018; 8(2): 555733. DOI: 10.19080/JFSCI.2018.08.555733

Future Recommendations

a) Text analysis of authors belonging to a particular gender can be done.

b) Whatsapp status updates can be used for the analysis.

References

1. Sounak Mitra (2014) Instant messaging apps' use doubled in India in last two years: study.
2. Chapter 21 Introduction to Instant Messaging Software.
3. http://www.nacada.ksu.edu/Resources/Clearinghouse/View-Articles/InstantMessaging.aspx.
4. Lyta Penna, Andrew Clark, George Mohay (2009) Challenges of Automating the Detection of Paedophile Activity on the Internet.
5. Dee Scrip (2016) Instant Messaging - Expressway for Identity Theft, Trojan Horses, Viruses, and Worms.
6. https://servicedesk.dc-uoit.ca/FAQ%20Sheets/cyber%20bullying.pdf.
7. Larry J Siegel Criminology: The Core (5th edn).
8. https://security.illinois.edu/content/instant-messaging-viruses.
9. Susan C Herring (2007) A Faceted Classification Scheme for Computer-Mediated Discourse Volume 4.
10. Carole E Chaski (2001) Empirical evaluations of language-based author identification techniques 8(1): 1350-1771.
11. Gerald R McMenamin (2002) Forensic Linguistics: Advances in Forensic Stylistics.
12. John Olsson (2004) Forensic Linguistics-An Introduction to Language, Crime and the Law. Language Arts & Disciplines.
13. Victoria Guillen Nieto, Chelo Vargas Sierra, María Pardiño Juan, Patricio Martínez Barco, Armando Suárez Cueto (2008) Exploring state of the art software for forensic authorship identification 8(1).
14. Andrea Nini (2014) Authorship profiling in a forensic context.
15. Colin Simon Michell (2013) Investigating the use of Forensic Stylistic and Stylometric Techniques in the analyses of Authorship on a publicly accessible social networking site (Facebook).

DOI : 10.19080/JFSCI.2018.08.555733