Authorship Authentication Using Short Messages from Social Networking Sites

Jenny S. Li, John V. Monaco, Li-Chiou Chen, and Charles C. Tappert
The Seidenberg School of Computing Science and Information Systems
Pace University, White Plains, NY 10606, USA
jennyli@us.ibm.com, vinmonaco@gmail.com, lchen@pace.edu, ctappert@pace.edu

***Abstract*—This paper presents and discusses several experiments in authorship authentication of short social network postings, an average of 20.6 words, from Facebook. The goal of this research is to determine the degree to which such postings can be authenticated as coming from the purported user and not from an intruder. Various sets of stylometry and ad hoc social networking features were developed to categorize short messages from thirty Facebook authors as authentic or non-authentic using Support Vector Machines. The challenges of applying traditional stylometry on short messages were discussed. The test results showed the impact of sample size, features, and user writing style on the effectiveness of authorship authentication, indicating varying degrees of success compared to previous studies in authorship authentication.**

Keywords- Authorship authentication; stylometry; social network; language-based security; intrusion detection

I. INTRODUCTION

Social network sites, such as Facebook, MySpace, or Twitter, attract billions of users [6]. While users may assume social networks provide a trusted environment for sharing information with friends and family, the information maintained by the social network sites could be compromised. For example, hackers could spam users with false messages, or hack into users' accounts and post fake messages on the users' behalves [11]. Authorship authentication is one of the trending security concerns in social networks. How can we tell if a message is posted by the real user and not by others disguised as the user?

This research investigates how we can authenticate a user by the way he/she writes in a short message posted on a social network. From the current state of the art, there is no authorship authentication mechanism built-in for any social networking site. Once a user is logged in, there is no re-authentication or detection of abnormal user behavior. If a hacker gains access to a user's account, he/she can disguise as the user – post messages, comment on the user's circle of friends' posts, or organize events on the user's behalf. The real user's friends may not suspect that the posts are not authored by their friend as each message posted is associated with the name of the user. As friends and family leave comments or share the fraudulent posts, the hacker can easily gain information from them.

One of the challenges of this research is to find a way to authenticate users' writing of relatively short messages in social networks, which tend to be much shorter than emails, blogs or regular articles. Because social network users can create many posts on a daily basis, it is not practical to apply the same kind of security features used on business transactions, such as Computers and Humans Apart (CAPTCHA), to social networking environments. Asking social network users to verify every post they make with such security features would create extra steps for them and reduce the usability of the social network. Hence, there is a need for non-obtrusive authorship authentication procedures for social network postings.

The objective of this research is to find an efficient and non-intrusive way to authenticate a user's post on a social network using the historical data maintained by the site. We used Facebook to illustrate the research methodology and experiments. We investigated this problem by revising traditional stylometry for long text to target Facebook posts which are much shorter.

Section 2 reviews literature on stylometry and authorship authentication. Section 3 describes our research methodology, data collection and processing. Section 4 describes our experimental design. Section 5 discusses our experiment results. Section 6 provides a conclusion.

II. LITERATURE REVIEW

Stylometry refers to study of linguistic style including sentence length, word choices, word count, syntactic structure etc. Stylometry reflects personal writing styles, defined in terms of stylometric features [18]. There are five common stylometric features [13] including *lexical features* such as character or word based, *syntactic features* such as use of function words ("and", "but", "on", etc.) or punctuation, *structure features* such as how the text is being laid out, *content-specific features* such as choice of words within a specific domain, and *idiosyncratic features* such as misspellings, grammatical mistakes, or deliberate author choices of words or cultural differences.

Stylometry is a well-established means of authorship authentication by using long text such as books, articles, etc. Researchers were able to achieve an accuracy rate of 70% to over 90%. Baayen et al. [12] tested 72 articles produced by 8 users with an average of 908 words per article. They achieved an accuracy of 88.1% by using 50 common function words and 8 punctuation symbols. Stamatatos 2007 [16] tested 100 messages that ranged from 288KB to 812 KB in size (1KB is roughly 500 words) with a modified Common N-Gram (CNG) method. They achieved an accuracy rate of about 70%. Koppel and Schler 2004 [3] used most frequent words to identify the author among 10 authors with 21 books of varying length per author. They achieved a 95.7% accuracy rate. Iqbal, et al. 2010 [18] used 292 stylometric features including lexical, syntactic, structural and topic specific to analyze the Enron Corpus with 158 users, each with 200 emails. They achieved an 82.9% accuracy rate**.**

Various researchers have investigated how effective was stylometry for authorship authentication using shorter text that ranged from 50 to a few hundreds of words. Their results were not as desirable as the results from the long text researches. Angela 2006 [15] created an instant message intrusion detection system framework using character frequency analysis to test 4 users' instant message conversation logs with 69 stylometric features including sentence structure, predefined specific characters, emoticons, abbreviations etc. The Naïve Bayes classifier produced the best accuracy with an average of 68% of data that lie within one standard deviation of either side of the mean. Corney, et al. 2002 [14] tested 4 users with 253 emails each that ranged from 50-200 words per email. They applied stylistic, structural and function words as measures and used SVM [6, 7] as the classification engine yielding 70.2% identification accuracy. Haytham Ohtasseb 2009 [17] investigated the methods for authorship authentication for online blogs/diaries. The researchers leveraged the LIWC (Linguistic Inquiry Word Count), MRC Psycholinguistic database, and a collection of syntactic features. They selected 63 LIWC features, tested 8 authors with 301 samples among them to achieve an average of 52.5% to 86% identification accuracy for blog lengths of 100 to 600 words. Alison and Guthrie 2008 [17] tested 9 users on short emails averaging 75 words each, with the number of emails per user ranging from 174 to 706, using 2-grams, 3-grams and word frequency measures. Using SVM as the classification engine, they achieved an average of 86.74% accuracy.

For extremely short text research, Layton et al. 2010 [10] tested 50 Twitter users with 120 tweets per user for authorship authentication. Twitter imposed messages to be 140 characters long for their maximum length (about 28 words). They used 3-gram and the SCAP classification method to obtain about 70% accuracy.

III. RESEARCH METHODOLOGY

We collected Facebook posts from 30 users including 6 friends who agreed to provide their posts and 24 public figures (such as movie stars, athletes, journalists, politicians etc.) that have their posts publicly accessible. To guarantee the confidentiality of these users, their identities would not be disclosed. Their data remained anonymous thorough the study. Close to 10,000 posts that were posted over the last four years were collected among these 30 users. The average number of posts per user was 308.6.

We created an AWK program to extract 233 features for each Facebook post and generated a feature file for each user. The features files were passed to the SVM Light [4, 5] program, an implementation of support vector machine (SVM) [6], as inputs for training and classification. For each user, we created an input file consisted of both positive data from the user and negative data selected randomly from others. Leave-One-Out (LOO) method was used for cross-validation. These training sets and testing sets for each user were fed into SVM Light for testing. We repeated the process for 30 users. False Acceptance Rate (FAR), False Rejection Rate (FRR), and accuracy rate were calculated for each user's test results:

FAR = No. of false acceptances/ No. of negative samples
FRR = No. of false rejections/No. of positive samples
Accuracy = 100% – (FAR+FRR)/2

The average result of the 30 users was calculated for each test case. The highest accuracy rate for each test case was also recorded. We repeated the process for all 12 test cases. We also performed 3 runs for each test case to obtain a more representative result. With the LOO method, we ran more than 666,000 tests for 12 test cases. See Section 4 for test results.

A. *Stylometry System*

We used 233 features in this study which included 227 stylometric features and 6 social network specific features. A portion of the 227 stylometric features in this research was selected from a subset of features from Zheng's research [1] to include character-based and word-based features. Other types of features from Zheng's such as structural features and content-specific features were not used as they were not applicable to the Facebook data. Zheng et al. studied authorship identification of online messages including messages from email, newsgroup or chat rooms. Compare to Zheng's studies, Facebook messages could be shorter than emails or newsgroup chats. The average length of our Facebook posts collected was 20.6 words or 103 characters assuming 5 characters per word. The length of our Facebook samples was compatible to the restricted character length of Twitter messages.

We added 6 social network specific features (features 228-233). These features included emoticons (a happy face and a sad face), abbreviation ("LOL"), starting a sentence without an uppercase letter, ending a sentence without a punctuation mark, and not mentioning "I" or "We" in the post. These features reflect a more causal writing style, which a user may not care about proper grammar or sentence structure. They write in a colloquial way that is similar to everyday conversation or a style that is commonly seen in chats or short text messages.

B. *Classification System*

We used SVM Light [4, 5], an implementation of Support Vector Machine (SVM) [6], as the machine learning and classification program. SVM Light provides four kernel functions: Linear, Polynomial, Gaussian radial basis function, and Sigmoid tanh. We tested a smaller sample size of 10 users' data with all of the functions. The default linear function produced the best result, hence, it was chosen.

Among the 30 users, some were less active. They had fewer than 50 posts over a few years. Using half of their data for training and half of it for testing was not practical and may not show a good representation of the users' profiles. To overcome this issue, we used the Leave-One-Out method [9] to maximize the sample size for training.

SVM Light allows customization of trade-off between training error and margin, which is represented in the C parameter [4]. The value of C provides flexibility to adjust the width of the soft margin from the hyperplane so that fewer training data falls on the wrong side of the hyperplane. To optimize the value of C, we performed a grid search of C from 0.01 to 2.0 with an incremental of 0.01. C=0.8 produced the best result which yielded a relatively close FAR and FRR for our samples.

IV. EXPERIMENTAL RESULT

A. *Impact of Features*

We conducted 12 sets of tests on Facebook data using the 233 selected features and SVM as the classifier. These tests aimed to discover the performance of stylometric and social network specific features, combined or separated, for authorship authentication.

TABLE I. AUTHORSHIP AUTHENTICATION TESTS WITH 233 FEATURES ON 30 USERS' FACEBOOK DATA

Test	**Features Tested**	**Accuracy Rate**	**FAR**	**FRR**	**Highest Accuracy Rate**	**Standard Deviation**
Test 1	All features (223 features)	79.6	14.3	26.5	95.2	6.7
Test 2	Stylometry only (227 features)	78.9	15.1	27	94.9	7.7
Test 3	Social network specific (6 features)	69.8	24.3	36	96.6	12.8
Test 4	Char based (50 features)	76	17.7	30.4	98.4	8.8
Test 5	Punctuations (8 features)	73.8	26.6	25.8	98.3	12.6
Test 6	Function words (150 features)	72.9	22.3	31.9	96.3	9.9
Test 7	No. of sentences (1 feature)	53.6	50.8	42.1	87.5	16.7
Test 8	Word based (8 features)	74.1	21.3	30.6	95.5	10
Test 9	Popular function words (33 features)	71.6	24.9	32	96.6	10.6
Test 10	Smilies (2 features)	67.8	54.4	10	99.6	18.1
Test 11	Missing upper case & period etc.(2 features)	67.9	28.2	36	98.6	15.9
Test 12	Not mentioning "I" & "We" (1 features)	60.8	40.5	38	98	18.4

Refer to Table I, the combined use of stylometric features and social network specific features (Test 1) produced the best accuracy rate of 79.6% among the 30 users. Stylometric features by themselves (Test 2) yielded a 78.9% accuracy rate, almost as good as the combination of stylometry and social network specific features together. This showed the selected 6 social network specific features provided a slight improvement to the overall accuracy when being combined with stylometric features.

In general, the six social special features alone (Test 3) were not as reliable. The list only yielded a 69.8% accuracy rate for the 30 users on average. However, we observed a phenomenon that social network specific features could be helpful in determining authorship if a user frequently used some of these features such as emoticons or others. The highest accuracy rate found among the 30 users for the 6 social network specific features was 96.6% vs. 95.3% when combined with stylometric features as in Test 1. This result was further supported by tests on individual social network specific features. The highest accuracy rate found among the 30 users for using only the smilies (Test 10) was 99.6%. It showed that one user used smilies on more than 82% of his/her posts while others rarely used smilies. Hence, this user's writing style was more distinctive. For Test 11, most users forgot to use uppercase to start a sentence occasionally. They could be careless, start a sentence with a hashtag by using the "#" sign or tag a person by using the "@" character, start a sentence with a quote from other people by using a quotation mark, etc. On the other hand, a lot of them have not used a proper punctuation to end a post. They may use too many punctuations such as "!!!" or "???", use different symbols to represent an emoticon or a face such as ">_<", ":-/", use character combinations to express feelings such as "xoxo" for hugs and kisses or "<3" for love, use a signature such as "-xxx" where "xxx" is the user's name or signature, post incomplete sentence as caption for a picture or link, etc. When a user wrote with proper capitalization for opening a sentence and used proper punctuation to close a sentence, the user's writing style would stand out from the rest. This was reflected with the highest accuracy rate of 98.6% from the result of Test 11. Test 12 investigated how often users talked about themselves by using "I" or "We" verses other topics. It was a surprise that a majority of our Facebook users talked about other topics for more than 60% in their posts. For users who talked about themselves most of the times, these users' writings styles were more distinctive. The highest accuracy rate was 98% for one of these users as he/she talked about him/herself for 88% of the time.

Most social network users in our study posted short messages. The average number of words per post was 20.6. Stylometric features that are character based (Test 4) would be more effective to determine authorship than word based features (Test 8). There may not be enough words for word based features to take effect for developing a user writing style profile. Among different types of stylometric features, character-based stylometric features (Test 4), alone yielded best accuracy rate of 76%, followed by word based features (Test 8), with a 74.1% accuracy rate, punctuation based features (Test 5) yielded a 73.8% accuracy rate, and function words (Test 6) generated a 72.9% accuracy rate. Sentence based feature (Test 7) showed the worst performance with a

53.6% accuracy rate. Since users tended to write short messages, most users would end up with one or two sentences. There was little to differentiate among messages by just looking at the number of sentences. The 150 function words were used in Test 6 including “about”, “from”, “if”, “and” “but” etc. As our Facebook messages were short, most likely that only a small subset of the function words was used in each post. We further selected a subset of 33 popular function words in Test 9. These words were used more than 10% of the total number of posts collected. The short list of function words yielded an accuracy rate of 71.6% while the full list of function words yielded 72.9% (in Test 6). More features being used would not harm the result of the tests. However, it costs more computational time to calculate values of more features. It would be a design decision of the social network authorship authentication provider to decide on the trade-offs between computational effort and accuracy.

B. *Impact of Number of Users*

We used all 233 features to test a batch of 10 users, 20 users and 30 users for the impact of number of users used (See Table III). Compare 10 users to 20 users, testing 10 users yielded an 81.6% accuracy rate while 20 users (that included the same 10 users as before) yielded a 79.8% accuracy rate. The tests of 30 users (that included the previous 20 users) yielded 79.6% accuracy rate. There was a slight advantage of using only 10 users as the accuracy rate was slightly better than 20 users or 30 users. However, the results between 20 users and 30 users were too close to conclude that the increase in number of users being tested decreased the accuracy rate in authorship authentication.

Users with distinctive writing styles were easier to be differentiated from the rest. In our tests (see Table II), we separated the users into 3 groups of 10 users. Both second and third groups contain users with very distinctive writing styles as showed by the highest accuracy rates of 95.3% and 94.9% for the second group and the third group respectively. The highest accuracy rate from the first group was only 85.4%. Therefore, a larger group of users with more distinctive writing styles can out perform a smaller group of users with less distinctive writing styles.

TABLE II. TESTING DIFFERENT USER GROUPS WITH 233 FEATURES

Test ID	Group of Users	Accuracy Rate	FAR	FRR	Highest Accuracy Rate
A	Group A: 1-10	77.9	21.7	22.4	85.4
B	Group B: 11-20	81.6	9.8	26.9	95.3
C	Group C: 21-30	79.3	11.4	30.1	94.9

TABLE III. TESTING DIFFERENT SIZES OF USER GROUPS WITH 233 FEATURES

Test ID	No. of Users	Accuracy Rate	FAR	FRR	Highest Accuracy Rate
Test 1a	10	81.6	9.8	26.9	95.3
Test 1b	20	79.8	15.8	24.6	95.3
Test 1c	30	79.6	14.3	26.5	95.3

C. *Impact of Number of Features*

Would more features clutter the analysis or decrease the accuracy rate? Our results from Table I showed a tendency that tests with more features showed a higher accuracy rate and a smaller standard deviation. Test 7 (number of sentences) and Test 12 (missing “I” and “We”) both had one feature. They produced the two worst results in terms of low accuracy rate and large standard deviation. Test 1 with all 233 features produced the best accuracy rate and had the smallest standard deviation. So far all of the tests results supported the argument that testing with more features would produce a better result except for the case for Test 4. Test 4 with 50 character-based features produced a better result (76% accuracy rate and 8.8% standard deviation) than Test 6 with 150 function words (72.9% accuracy rate and 9.9% standard deviation). This simply confirmed that character-based features were more desirable measures for short text authorship authentication than word based features.

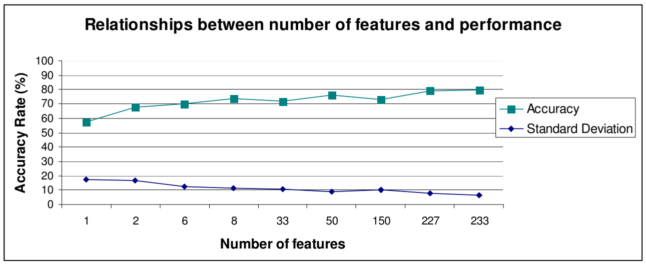

Figure 1. Impact of number of features

D. *Tesing with the k-NN Algorithm*

To compare with another algorithm the data were re-tested using the k-Nearest-Neighbor algorithm. It uses Euclidean distance to classify the unknown difference vectors, with a reference set composed of the differences between all combinations of the claimed user’s enrolled vector (within-person) and the differences between the claimed user and every other user (between-person). The differences of difference vectors are being calculated [8].

Using the k-Nearest-Neighbor method, the average accuracy among the 30 users for Test 1 (with all 233 features) was 65.5%. The result showed that SVM yielded a much better result of 79.6% accuracy rate (see Table I Test 1) than the k-Nearest-Neighbor algorithm for short messages.

E. *Tesing with Normalized Stylometric Features*

The list of selected 233 combined stylometric features and social network specific features described above was not normalized. We basically counted the frequency of each feature as it appeared in each post. We used the stylometric features that were used in Monaco et al.’s research [8] for re-testing. The goal was to investigate if another set of stylometric features would yield better results on our

Facebook data. Monaco's features were normalized, which represented the ratio of each feature against the whole message. Monaco used his set of 228 stylometric features for authorship identification of 30 book authors. Each had 10 book samples; each book was about 10000 words or longer.

Using Facebook data, Monaco's stylometric features, and SVM, the test result was 77% accuracy with 17.4 FAR and 28.7 FRR. This result was very closed to the 227 stylometric features test from Test 2 (See Table I Test 2) that yielded 78.9% accuracy with 15.1% FAR and 27% FRR. Even though Monaco's features were not exactly the same as our features, there were duplications including the character-based, word-based and some syntax-based features. It was hard to conclude if normalized or un-normalized features were more applicable for testing with short text. Both results were similar although un-normalized features showed slight improvement in performance.

F. *Contributions and Limitations*

Our research was the first study that investigated short message authentication using Facebook data, which was much shorter than emails, blogs, articles or books. Classifiers that typically work well for long text might not work for short messages and specific features might be needed. This new research problem requires novelty in selection of features, classifiers, and the sensitivity of the feature selection and classifier selection.

We used SVM as the classifier, linear kernel function and soft margin optimization on training data. Layton's Twitter research [10], 140 characters per tweet, was the only research we were aware of that analyzed extremely short messages. They yielded 70% accuracy rate with the 3-gram and the SCAP classification method. Our research on Facebook data (with 20.6 words or 103 characters on average per post) achieved 79.6% accuracy rate with SVM and a combination of stylometric and social network specific features.

Zheng's research [1] focused on email and newsgroup posts. Monaco's research [8] focused on novels. Our average Facebook data message length was far shorter than emails, newsgroup posts or novels, which put our research in a different domain. We did fuse both Zheng's features and Monaco's features with our Facebook specific features. We showed the impact of different feature combinations and provided insights into the sensitivity of various feature sets on the accuracy rate; see Table I. This will be valuable for future research in new social media where messages are often short in nature. Character-based features are more applicable than word-based features for social network posts. This aligns with the fact that social network posts are much shorter.

We identified 6 social network specific features that could be useful for improving authorship authentication of messages posted in a social network. These features showed an accuracy rate as high as 96.6% (Test 3) for users who adopted them extensively in their writing style.

While these social network specific features can potentially yield very high accuracy rate for certain users, it is also a limitation of our work. From our results, extensive use of smilies can generate as high as 99.6% accuracy. Hackers or unauthorized users can spend time to study the writing style of the person they would like to mimic. We do not recommend using these features alone. It is more reliable to combine these social network specific features with some stylometric features for more reliable results.

We have tested the same Facebook data with two different classification algorithms: SVM with a linear kernel, and k-Nearest-Neighbor. Our tests showed that SVM was a better classification algorithm for our data set.

V. CONCLUSION

While social networks have gained tremendous popularity, they have also created security threats to users [2]. Spam, flaw in third-party applications, worm, phishing are just some sample attack methods that hackers can use to gain information from others. This research investigated authorship authentication of Facebook postings, a fundamental trust concern of whether messages are posted by legitimate users or not. Given a set of available messages posted by a user, this study aimed to determine if a new and possibly disputed message is authored by the same user. We faced two challenges. The first challenge was that social network messages tend to be much shorter than novels, blogs, or emails, and our concern was whether traditional stylometric features would be effective in authorship authentication for these short messages. The second challenge was the limited number of posts from some users. Because some users post infrequently it is not reasonable to divide these users' data into halves for training and testing. We overcame these challenges. We offered a solution with a combination of traditional stylometric features and social network specific features as measures, SVM as the classifier with a linear kernel function and optimization of the C parameter that managed the width of soft-margin from a hyperplane. The Leave-One-Out method was used to accommodate the limited amount of posts collected.

To our knowledge this study was the first to investigate authorship authentication on Facebook posts. Test results of using all 233 combined stylometric and social network specific features showed an accuracy rate of 79.6% when verifying whether a message was written by the real user. It provided a prediction to questions such as "Does this message look like a message that would be written by the user?", "Can we trust this message?" We tested individual subsets of our features and showed the impact of different feature sets. Our study gives social network providers an overview of the trade-offs in case they are interested in building an authorship authentication solution to protect their users' accounts. In the future, this research can be extended to use the same list of features for testing long messages such as blog, emails, novels etc. By doing so, we can evaluate the effectiveness of these features in long messages verses short messages for authorship authentication. We are in the process of re-testing our test cases with more classifiers for effectiveness on short text. More experiments are underway to explore areas for improvements.

REFERENCES

[1] R. Zheng, J. Li, H. Chen, and Z. Huang., "A framework for authorship identification of online messages: Writing-style features and classification techniques." Journal of the American Society for Information Science and Technology 57(3): 378-393. 2006

[2] Weimin Luo, Jingbo Liu, Jing Liu, Chengyu Fan, “An Analysis of Security in Social Networks”, Proceeding of the Eighth IEEE International Conference on Dependable, Automatic and Secure Computing, 2009.

[3] M. Koppel and J. Schler, “Authorship verification as a one-class classification problem”, ICML ’04 Proceedings of 21st International Conference of Machine Learning, New York, 2004.

[4] Thorsten Joachims, Learning to Classify Text Using Support Vector Machines. Kluwer Academic Publishers, 2002.

[5] SVM-Light Support Vector Machine, http://svmlight.joachims.org/

[6] C. Cortes, V. Vapnik, Support Vector Networks. “Machine Learning”, 20:273-279, 1995.

[7] V. Vapnik, “The Nature of Statistical Learning Theory”, Springer-Verlag, New York, 1995.

[8] J. Monaco, J. Stewart, SH Cha, C. Tappert, “Behavioral Biometric Verification of Student Identity in Online Course Assessment and Authentication of Authors in Literary Works”, IEEE 6th International Conference on Biometric, BTAS 2013

[9] M. Kearns, D. Ron, “Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation”, MIT Press Journals, 1997

[10] R. Layton, P. Watters, and R. Dazeley, “Authorship attribution for twitter in 140 characters or less”, Second Cybercrime and Trustworthy Comp. Workshop, 1-8, 2010.

[11] “Facebook protects users following Adobe hack attack,” BBC Technology News. 13th, November, 2013. Available at http://www.bbc.co.uk/news/technology-24925874. Accessed at 11th, November, 2013.

[12] H. Baayen, H. van Halteren, A. Neijt, and F. Tweedie, “An experiment in authorship attribution”, 6th JADT, 2002.

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

[14] M. Corney, O. de Vel, A. Anderson, and G. Mohay, “Gender-preferential text mining of e-mail disclosure”, Proceeding of 18th Annual Computer Security Application Conference, Las Vegas, NV, Dec 2002.

[15] O. Angela, “An Instant Messaging Intrusion Detection System Framework: Using character frequency analysis for authorship identification and validation. Carnahan Conferences Security Technology”, Proceedings 2006 40th Annual IEEE International. 2006.

[16] E. Stamatatos. “Author Identification Using Imbalanced and Limited Training Texts”. 18th International Workshop on Database and Expert Systems Applications, 2007.

[17] B. Allison and L. Guthrie, “Authorship attribution of e-mail comparing classifiers over a new corpus for evaluation”, Processing of LREC ’08, 2008.

[18] F. Iqbal., L. A. Khan, et al., ”e-mail authorship verification for forensic investigation”, Proceedings of the 2010 ACM Symposium on Applied Computing, Sierre, Switzerland, ACM: 1591-1598, 2010.

APPENDIX A – LIST OF FEATURES USED

Character-based features:
Feature 1: number of characters
Feature 2: number of alphabets
Feature 3: number of uppercase characters
Feature 4-29: number of alphabet a-z
Feature 30-50: number of special character “~ @ # $ % ^ & * - _ = + > < [ ] { } / \\ |"

Syntactic Features:
Feature 51-58: number of punctuation “, . ? ! : ; \" ' "
Feature 59-208: Function words
“a, about, above, after, all, although, am, among, an, and, another, any, anybody, anyone, anything, are, around, as, at, be, because, before, behind, below, beside, between, both, but, by, can, cos, do, down, each, either, enough, every, everybody, everyone, everything, few, following, for, from, have, he, her, him, I, if, in, including, inside, into, is, it, its, latter, less, like, little, lots, many, me, more, most, much, my, need, neither, no, nobody, none, nor, nothing, of, off, on, once, one, onto, opposite, or, our, outside, over, own, past, per, plenty, plus, regarding, same, several, she, should, since, so, some, somebody, someone, something, such, than, that, he, their, them, these, they, this, those, though, through, till, to, toward, towards, under, unless, unlike, until, up, upon, us, used, via, we, what, whatever, when, where, whether, which, while, who, whoever, whom, whose, will, with, within, without, worth, would, yes, you, your”

Structural Features:
Feature 209: Total number of sentences

Word-based features:
Feature 210: Total number of words
Feature 211: Total number of short words (less than four characters)
Feature 212: Average word length
Feature 213: Average sentence length in terms of character
Feature 214: Average sentence length in terms of word
Feature 215: Number of words with 1 char
Feature 216: Number of words with 2 chars
Feature 217: Number of words with 3 chars
Feature 218: Number of words with 4 chars
Feature 219: Number of words with 5 chars
Feature 220: Number of words with 6 chars
Feature 221: Number of words with 7 chars
Feature 222: Number of words with 8 chars
Feature 223: Number of words with 9 chars
Feature 224: Number of words with 10 chars
Feature 225: Number of words with 11 chars
Features 226: Number of words with 12 chars
Features 227: Number of words with more than 12 chars

Social Network Specific features:
Feature 228: Frequency of a happy face “:)”
Feature 229: Frequency of a sad face “:(“
Feature 230: Frequency of “LOL”
Feature 231: Frequency of missing an uppercase letter when starting a sentence
Feature 232: Frequency of missing a period or other punctuation to end a sentence
Feature 233: Frequency of missing the word “I” or “We” in a sentence