Annales Mathematicae et Informaticae
56 (2022) pp. 10–21
DOI: https://doi.org/10.33039/ami.2022.12.005
URL: https://ami.uni-eszterhazy.hu

A psychometric approach to email authorship assertion in an organization

Prathamesh Berde[a], **Manoj Kumar**[b], **C.S.R.C. Murthy**[b], **Lalit Dagre**[b], **Seervi Tejaram**[b]

[a]Homi Bhabha National Institute Mumbai, India
prathameshb@hbni.ac.in

[b]Bhabha Atomic Research Centre, Mumbai India
kmanoj@barc.gov.in
murthy@barc.gov.in
lalitd@barc.gov.in
tejas@barc.gov.in

Abstract. Email services have become an integral aspect of modern communication. Emails can be transmitted digitally without the adequate authentication of the sender. As a result, there has been a considerable surge in security threats coming from email communication, such as phishing, spear phishing, whaling, and malware deposition through emails where recipients can be duped into acting. Authorship assertion of the sender can prevent several security issues, particularly in an organizational setting where an employee's trust can be compromised by faking an email from a colleague or senior without exposing any specific system weakness. A psychometric approach to determining the authorship of an email in an organization is proposed in this research. Machine learning (ML) models have been developed using four classification algorithms. The performance of these ML models has been compared.

Keywords: authorship, personality, machine learning, psychometric features

1. Introduction

The Internet has become an integral part of our life. In modern-day communication, the predominant mode of communication on the internet is Email. Email service impales very deep into private networks and intranet of organizations, thereby allowing attackers to deploy the exploits far into organizations' networks. Hence,

Submitted: October 26, 2022
Accepted: December 3, 2022
Published online: December 28, 2022

the security of email service is one of the major tasks in an organization. One of the prominent attacks on email is the social engineering attack. The knack of influencing the people to divulge sensitive information of some other action is known as social engineering and the process of doing it is called the social engineering attack [13]. In some of the modern-day social engineering attacks against one victim or a small group thereof, the attackers research their targets to design phishing emails specific for each victim. The emails appear to be coming from a trusted colleague/party and prompt the recipient to follow the directions inside. By impersonating trusted email senders through meticulously crafted messages, attackers trick the receivers to act on that email and launch malware. Such an attack is mostly used as a platform for injecting malware into interior parts of an organization such as the Intranet. Attacks involve targeting individuals from organizations by maneuvering them to promulgate misleading information to varied interests and valuable and sensitive data that may intrigue cybercriminals without exploiting a specific vulnerability. As discussed in [1], emails can transmit information digitally without authenticating the person who writes the text and could be used by criminals for malicious intentions. Authorship assertion of such emails becomes necessary in an organization.

Alhijawi et al. [1] surveyed some of the possible techniques for authorship attribution. They carried out the authorship analysis technique to satisfy the objective. Authorship Identification, similarity detection, and characterization were its three main perspectives. Their survey showed the use of stylometric features for authorship identification. The features were classified into four categories namely lexical, character, syntactic and semantic. Lexical features included token-based, vocabulary richness, word frequencies, word n-grams, errors, character features included character types, character n-grams, etc. Syntactic and semantic features included the parts of speech and semantic dependencies. Some of the datasets in the research were email datasets, online text data sets, source code data sets, etc. Yet, it is observed that the features used in this research may not be invariant as the context of the writing changes.

One of the approaches in this field is the classification of authors' emails based on their representation of text to vectors [4]. Here, they used the word2vec to generate the word embedding and extract the features of the author's writing style from their text writing. Multi-layer Perceptron classifier and the back-propagation learning algorithm were used for classification. They used the PAN12 free fiction collection data corpus written in English. A cluster-based classification model for email authorship identification was also used [15]. Stylometric features like punctuation used at the end of the emails, the tendency of the user to start the emails with the capitalized letters, punctuation after the greetings and farewells, etc were used for classification. The dataset used for their analysis was the Enron email dataset.

One of the other works in this field, carried the authorship identification for short online messages [5] using Supervised Learning and n-gram analysis. Enron email dataset was used for their analysis. One of the works used an approach of

Unsupervised Clustering for authorship identification [14] for email forensics where they classified emails initially using unsupervised clustering and then identified the stylometric features in the clusters. They used the Hierarchical Clustering and Multidimensional scaling approach of Unsupervised Clustering for authorship identification. They also used the Enron email corpus data set for their experimentation.

The motive behind carrying out the work presented in our paper was to develop classification models of known authors in an organization so that the impersonated emails claiming to be coming from these authors could be asserted. Hence, this work emphasized developing models that assert authorship of an email in an organization using Machine Learning algorithms for known email authors.

The remainder of the paper is organized as follows. Section 2 introduces the methodology used for authorship assertion. Section 3 presents the details of feature extraction and training of the ML classifier. Validation of feature extraction models is discussed in Section 4. An analysis and comparison of performance metrics of different ML models are discussed in Section 5.

2. Methodology

The proposed approach to email authorship assertion in this paper is based on the fact that personality is a stable and invariant aspect of an individual [9] and the most relevant differences/traits are encoded in the language written [3]. Using these characteristics of the personality and language (extracted from emails), the problem of authorship assertion is transformed into a classification problem. To formulate the classifier, the following are needed:

2.1. Evaluation of personality score from the questionnaire

Personality is the characteristic pattern of those sensory, perceptual and cognitive systems within an individual that determines his unique behavior in his environment [2]. The Big Five Personality Model is one of the most widely used models of personality. This model is also known as the five-factor model or the OCEAN model which is based on five personality dimensions i.e. Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism [9]. Volunteering authors undergo a personality assessment test and personality scores are generated. The scores are based on the International Personality Item Pool proxy for the NEO Personality Inventory-Revised (NEO PI-R) questionnaire [6]. NEO PI-R is considered by many psychologists for measuring the dimensions within the Big Five Personality Model.

Statistics about the personality dimensions evaluated from the questionnaire have been given in Table 1.

Table 1. Statistics of personality scores of users after NEO PI R personality questionnaire.

Dimension	**Mean**	**Standard Deviation**
Neuroticism	48.17	30.41
Openness	30.17	23.75
Agreeableness	62.53	22.23
Extroversion	43.74	28.56
Conscientiousness	67.03	21.13

2.2. Extraction of word category lexica from emails

Various word categories are described in the word category lexica of the content-coded dictionary of the packages provided in [7, 20] available on LIWC [17]. Word count corresponding to various parts of speech (POS) categories like articles, conjunctions, etc. using Spacy [10] in the Python programming language is extracted from the emails. The word count corresponding to each word category like positive, negative words, sadness, achievements, etc. in the dictionary using the Empath [7] package in the Python programming language is derived from the emails. The word count corresponding to each category is appended to a column vector for an email.

2.3. Feature vector extraction for classifier and authorship assertion

The feature vector for the classifier consists of a score of personality corresponding to the personality dimension in the five-factor model. To extract the personality dimension scores, a regression model may be used. The regression model estimates the personality score using the correlation of personality score evaluated from the authors' questionnaire and their corresponding emails' column vectors as discussed in Section 2.2. The classifiers are trained using features of old emails and subsequently used for authorship assertion of new emails they claim to be coming from.

For implementing regression models to extract personality scores, Linear Regression, Support Vector Regression (SVR), Regression Trees, and Neural Networks have been used. For the classification of emails in the last stage, Logistic regression, Support Vector Machine (SVM), Neural Networks, and Naive Bayes have been used. All the algorithms have been implemented in Python 3 using the modules of Scikit-learn [16].

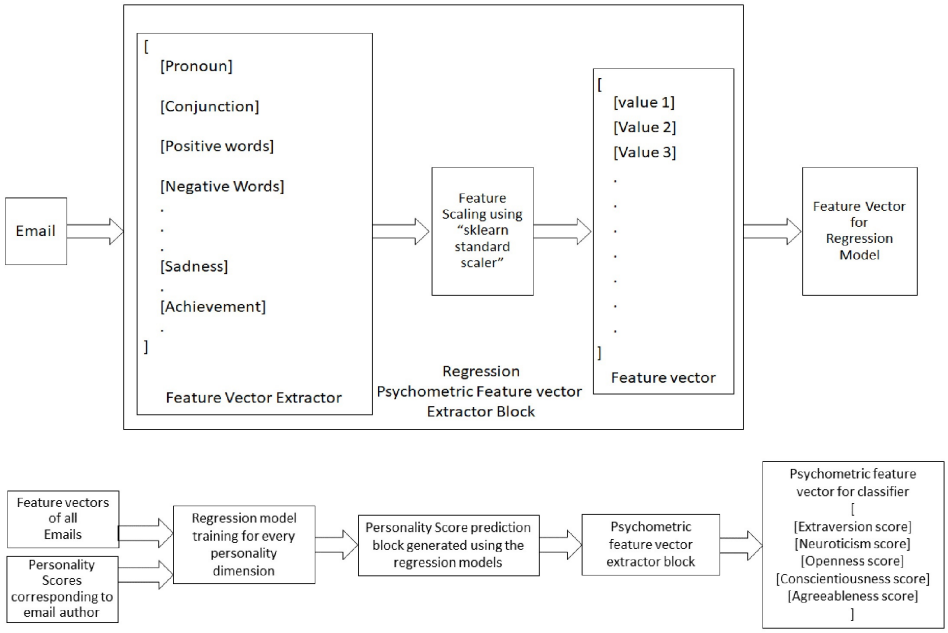

Figure 1. Psychometric feature vector extraction.

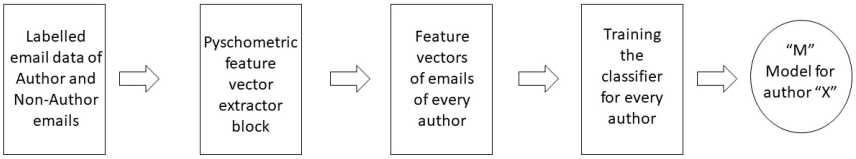

Figure 2. Training of classifier using Psychometric Features.

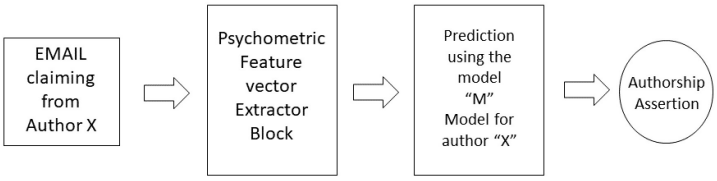

Figure 3. Classification Stage.

3. Implementation

For authorship assertion in the organization, the experiment was conducted on a limited set of 18 users. These users have volunteered, given consent to use their

past emails, and answered the questionnaire for personality dimensions [6]. We tried to develop the author-specific models to analyze if the email had been sent by him or not. Out of all the authors who volunteered, the classifier analysis of the 5 authors who had the highest number of emails is discussed in this paper.

3.1. Data preparation and pre-processing

The first stage of the implementation was data preparation. In the data preparation, a data frame was prepared for analysis of the data. The sent emails of the authors were used. The sent emails of the users had been collected for the past 1 year and only those emails were considered in which the author had started the conversation. The forwarded and replied emails were not considered in the analysis. Using the standard python programming language libraries, we pre-process the data and extract the text corresponding to the email bodies. The email body content for every email was separated after extracting the message in the email and the signatures from the emails were stripped off as discussed in [8, 21]. Emails were appended in a data frame. Now corresponding to every processed email, the score of personality dimension which had been collected from the questionnaire of the corresponding user was assigned.

3.2. Feature extraction and training

Regression techniques were used to relate word categories with authors' personality scores. As shown in Fig. 1, the scores for each personality dimension of the author were assigned and the counts corresponding to each lexical category of word category lexica were extracted from emails as inputs to the regression model as discussed in Section 2.2. Regression algorithms were used to fit a curve between independent factors i.e. the lexical categories and the regressand i.e. Extraversion, Neuroticism, Openness, Agreeableness, and Conscientiousness. The following steps were involved.

- Features were extracted by obtaining word count corresponding to various parts of speech and the word count corresponding to every lexical category for every email in the dataset using respective packages in python programming language as discussed in Section 2.2.
- Feature scaling was performed as the features varied in terms of what they represent. Some algorithms are invariant to feature scaling while some are not.
- Once the features were scaled, the regression models were trained using the regression algorithms specified in Section 2.3.
- Regression Algorithms like SVR and Neural Networks used various hyperparameters while training. Optimum hyperparameters for improving performance were chosen by hyperparameter tuning.

- After the Hyperparameters had been optimized, results and performance of the machine learning algorithms were compared and the model with the best performance evaluated using standard metrics[11] in Regression was chosen for the prediction of the score for every personality dimension.

- To verify whether the regression model correctly prepared data for the classifier and whether the features used for machine learning were sufficient to be used for a classifier, clustering analysis using K-means clustering and the Density-based spatial clustering of applications with noise (DBSCAN) clustering algorithm was performed and the goodness of the cluster was analyzed using standard metrics[18, 19].

3.3. Email classification for authorship assertion

After the generation of the regression model, a data set of emails was prepared for every author. In a particular data set, we selected all the emails which belonged to that author, and we randomly selected an equal number of emails that did not belong to that author. Then, for every email, we extracted the personality scores using the regression model generated in the previous step. As shown in Fig. 2, we extracted the feature vectors and modeled the author-specific classifier. The following steps were involved in this stage.

- An author and his email data for a year were collected. Then, we randomly selected the same number of emails as the author from data that did not belong to the selected author and labeled them correctly.

- The personality scores to each of these emails from the dataset were extracted using the regression models for each dimension in the previous stage and a feature vector matrix was derived which was followed by feature scaling.

- Once the features were scaled, the classification models were trained using the classification algorithms specified in Section 2.3 and the hyperparameters were optimized.

- After the Hyperparameters were optimized, we compared the results and performance of the machine learning algorithms and chose the algorithm with the best performance after analyzing using standard metrics for classification[11], and saved the model for use in the classification of email whether it belongs to the specified author.

As shown in Fig. 3, when a new email was received we first extracted the features using the regression model i.e the personality scores, prepared the feature vector matrix, and then predicted the class of this vector using the classification model of the author it claims to be coming from.

4. Validation of regression models

The performance of regression algorithms is given in Table 2. It is evident from the results that SVR outperformed any other regression algorithm for this data. It was also evident in the literature survey that kernelized regression algorithms like SVR have performed better than other algorithms. R^2 value for SVR was higher than other regression algorithms. The Mean Absolute Error (MAE) percentage was also relatively less compared to other regression algorithms. The decision to use these metrics for selecting the models was based on the facts published in the literature survey [12]. It is also to be noted that R^2 values mean the percent of explained variance on the dependant variable. So in our experiments when we tried to analyze the impact of a certain limited number of variables on the human-related outcomes, it was very difficult to explain the majority of the variance.

Table 2. Performance of personality score prediction model for psychometric feature vector extraction.

	Neuroticism		**Openness**		**Agreeableness**		**Extraversion**		**Conscientiousness**	
Algorithm	R^2	% MAE	R^2	% MAE	R^2	% MAE	R^2	% MAE	R^2	% MAE
Linear regression	0.15	34.37	0.15	25.04	0.3	18.34	0.14	18.83	0.15	30.37
Support vector Regression	0.43	12.65	0.36	13.47	0.41	13.19	0.44	12.32	0.42	15.96
Decision tree regression	0.29	23.94	0.18	18.63	0.24	23.86	0.31	15.25	0.28	21.43
Neural Network regression	0.18	27.3	0.16	23.67	0.35	17.28	0.24	16.11	0.19	26.28

Table 3. Performance of clustering algorithms to analyze data separability.

	K-means Clustering		**DBSCAN Clustering**	
No. of users	3 users	5 users	3 users	5 users
Estimated clusters	-	-	4	6
Silhouette Coefficient	0.714	0.69	0.68	0.59
Homogeneity	0.886	0.782	0.77	0.697
Completeness	0.881	0.78	0.774	0.63
V- Measure	0.884	0.781	0.771	0.662

To verify whether the regression model correctly prepared data for the classifier and whether the features used for machine learning are sufficient to be used for a classifier we performed clustering analysis using K-means clustering and the DBSCAN clustering algorithm. To perform the clustering analysis, 5 volunteers out of 18 having the highest number of emails were considered. It was evident from the results shown in Table 3 of the clustering analysis that the SVM regression

model, has features sufficient to explain the variation in personality and can be used to derive the features for training the classifier and we can model a supervised classifier for the analysis of the same.

5. Results and discussion

Metrics like accuracy, f-score, sensitivity, specificity, training time, and prediction time were evaluated for the choice of the best models. It was desired that emails that do not appear to be coming from the author should be asserted correctly as such emails may create havoc if undetected. We chose to decide on the best model by comparing prediction accuracy, prediction time, and specificity. In this work, we were able to achieve accuracy which was in the range of 80-95% for authorship assertion. The features used relied on the personality dimensions of the five-factor model of personality. It was observed from the performance of classification algorithms shown in Table 4 that the Neural Network classifier and the SVM classifier have comparable performance considering the accuracy of the model trained using psychometric features. These two classification algorithms perform better than Naive Bayesian and the Logistic Regression classification algorithm. From the clustering analysis, we observed that although the data used for training the classifier was separable, it is not perfectly homogeneous i.e. each cluster did not have data points belonging to the same class label. The SVM algorithm implemented in the classifier used in this approach required two hyperparameters, C and γ along with kernel functions to separate the two classes using a hyperplane. Kernel functions only calculated the relationship between every pair of points as they are in a higher dimension. Parameter C traded off misclassification of training data points against decision surface while γ determined how much influence a single training datapoint has. Optimum choice of the kernel function, values of C and γ were predicted using hyperparameter tuning.

The neural network learned the nonlinear function approximator using the summation of weighted layers of neurons and their transformation at the output of each neuron using its activation function for two classes using the various hyperparameters and optimum hyperparameters were obtained by hyperparameter tuning. Neural networks required a higher training time as the initialization of weights was done according to standard method i.e. by initializing weights and bias of the complex neural network by random number generation and were optimized by error backpropagation using stochastic gradient descent solver after every iteration, although the prediction time was not much higher as the weights had been tuned during the training phase. Due to the above reasons, SVM and Neural Networks were able to fit and perform better than other algorithms on the nonlinear and not perfectly homogeneous data points used in this analysis.

In the training phase as well as the testing phase, no other classifier was as fast as the Naive Bayesian classifier (the value of this metric was 2-3 milliseconds) because training the Naive Bayes classifier required the calculation of the probability of individual classes and the class conditional probabilities. Also, optimization pro-

Table 4. Performance of classification algorithms.

user	algorithm	accuracy	f1_score	sensitivity	specificity	training time (in s)	prediction time (in ms)
USER 1	Logistic regression classifier	86.86	0.87	0.87	0.9	0.015	0.002
	SVM classifier	90.06	0.9	0.9	1	0.345	0.072
	Neural Network classifier	89.74	0.9	0.9	0.95	1.749	0.005
	Naive Bayes classifier	88.78	0.89	0.89	0.91	0.003	0.003
USER 2	Logistic regression classifier	86.33	0.86	0.86	0.83	0.02	0.003
	SVM classifier	89.45	0.89	0.91	0.97	0.362	0.079
	Neural Network classifier	94.92	0.95	0.95	0.96	0.869	0.005
	Naive Bayes classifier	90.63	0.9	0.89	0.82	0.003	0.003
USER 3	Logistic regression classifier	94.32	0.94	0.94	0.89	0.025	0.002
	SVM classifier	95.63	0.96	0.95	0.92	0.127	0.036
	Neural Network classifier	94.76	0.95	0.94	0.9	0.605	0.005
	Naive Bayes classifier	95.63	0.96	0.95	0.92	0.003	0.003
USER 4	Logistic regression classifier	80.37	0.8	0.81	0.78	0.02	0.003
	SVM classifier	90.8	0.9	0.89	1	0.113	0.047
	Neural Network classifier	85.28	0.85	0.85	0.86	0.633	0.006
	Naive Bayes classifier	85.89	0.86	0.86	0.87	0.002	0.003
USER 5	Logistic regression classifier	84.81	0.85	0.85	0.86	0.02	0.003
	SVM classifier	87.97	0.88	0.87	0.99	0.114	0.047
	Neural Network classifier	92.41	0.92	0.92	0.99	0.625	0.006
	Naive Bayes classifier	82.91	0.83	0.84	0.73	0.002	0.003

cedures did not require the calculation of coefficients. Additionally, the algorithm assumes all features to be independent, and hence parametric calculations can be done individually and faster.

The prediction using SVM is comparatively slower because before prediction SVM transforms the input vector to a higher dimensional feature vector. Additionally, SVM used kernel trick to reduce the computation time in high dimensional feature space. Prediction time using all the algorithms is comparable in a few microseconds. Another important aspect that we analyzed was specificity. Specificity determined the fraction of actual negative cases which got predicted correctly. In our data, actual negative cases were those emails that do not belong to that user. We observed that the SVM classifier outperformed other classifiers on this metric (the value of this metric existed between 0.9 and 1). Hence, the use of an SVM classifier to train the classification model using the psychometric features is recommended.

6. Conclusion

The proposed technique is based on the fact that a person's personality is a constant and stable quality that is represented in his language. The authorship assertion problem has been treated as a classification problem using these principles. To develop the classifier, a questionnaire to assess personality traits has been used, then

the extracted word category lexica from emails are used to develop the personality score prediction model, followed by feature vector extraction and training of classifiers. A comparison of models developed using four classification algorithms was conducted to evaluate and choose the best model for each author based on parameters like accuracy, specificity, prediction time, and so on. On these metrics, SVM and Neural Network classifiers outperformed others.

Although these models function commendably, there may be inconsistencies if the threat actor and the real sender have similar personalities. Another inconsistency may develop if the personality scores collected via the personality questionnaire have not been attempted truthfully, since this may represent misleading personality behavior in the scores, making the training of the regression model erroneous. The work can be improved in the future by defining a more comprehensive set of features and employing advanced machine learning models. Model boosting and bagging may also increase performance and the development of models.

Acknowledgement. We would like to express our sincere gratitude to the Head, Computer Division, BARC for providing us with the data. We would thank Shri Rohitashva Sharma for providing the necessary infrastructure and allowing us to carry out this work at HBNI Complex. We would also thank Shri Shankar for the support.

References

[1] B. Alhijawi, S. Hriez, A. Awajan: *Text-based Authorship Identification - A survey*, in: 2018 Fifth International Symposium on Innovation in Information and Communication Technology (ISIICT), 2018, 1–7.

[2] G. W. Allport: *Personality: A Psychological Interpretation.* (1937).

[3] G. W. Allport, H. S. Odbert: *Trait-names: A Psycho-Lexical Study.* Psychological monographs 47.1 (1936), p. i.

[4] N. E. Benzebouchi, N. Azizi, N. E. Hammami, D. Schwab, M. C. E. Khelaifia, M. Aldwairi: *Authors' Writing Styles Based Authorship Identification System Using the Text Representation Vector*, in: 2019 16th International Multi-Conference on Systems, Signals Devices (SSD), 2019, 371–376.

[5] M. L. Brocardo, I. Traore, S. Saad, I. Woungang: *Authorship Verification for Short Messages using Stylometry*, in: 2013 International Conference on Computer, Information and Telecommunication Systems (CITS), IEEE, 2013, 1–6.

[6] P. T. Costa Jr, R. R. McCrae: *The Revised NEO Personality Inventory (NEO-PI-R).* Sage Publications, Inc, 2008.

[7] E. Fast, B. Chen, M. S. Bernstein: *Empath: Understanding topic signals in large-scale text*, in: Proceedings of the 2016 CHI conference on human factors in computing systems, 2016, pp. 4647–4657.

[8] H. Gascon, S. Ullrich, B. Stritter, K. Rieck: *Reading Between the Lines: Content-Agnostic Detection of Spear-Phishing Emails*, in: Research in Attacks, Intrusions, and Defenses, Springer International Publishing, Springer, Cham, 2018, 69–91, isbn: 978-3-030-00470-5.

[9] L. R. Goldberg: *An Alternative "Description of Personality": The Big-Five factor structure.* Journal of Personality and Social Psychology 59.6 (1990), p. 1216.

[10] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd: *spaCy: Industrial-strength Natural Language Processing in Python*, https://doi.org/10.5281/zenodo.1212303, 2020, doi: {10.5281/zenodo.1212303}.

[11] Joshi, Ameet V: *Machine Learning and Artificial Intelligence*, Springer, 2020.

[12] F. Mairesse, M. A. Walker, M. R. Mehl, R. K. Moore: *Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text.* Journal of Artificial Intelligence Research 30 (2007), 457–500.

[13] F. Mouton, L. Leenen, H. S. Venter: *Social Engineering Attack Examples, Templates and Scenarios*, Computers & Security 59 (2016), pp. 186–209.

[14] S. Nirkhi, R. Dharaskar, V. Thakare: *Authorship Verification of Online Messages for Forensic Investigation*, Procedia Computer Science 78 (2016), 1st International Conference on Information Security & Privacy 2015, 640–645, issn: 1877-0509, doi: {https://doi.org/10.1016/j.procs.2016.02.111}, url: %7Bhttp://www.sciencedirect.com/science/article/pii/S1877050916001137%7D.

[15] S. Nizamani, N. Memon: *CEAI: CCM-based email authorship identification model*, Egyptian Informatics Journal 14.3 (2013), pp. 239–249, issn: 1110-8665, doi: https://doi.org/10.1016/j.eij.2013.10.001, url: http://www.sciencedirect.com/science/article/pii/S1110866513000039X.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay: *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research 12 (2011), pp. 2825–2830.

[17] J. W. Pennebaker, R. L. Boyd, K. Jordan, K. Blackburn: *The Development and Psychometric Properties of LIWC2015*, http://liwc.app/, 2015.

[18] Rosenberg, Andrew and Hirschberg, Julia: *V-measure: A Conditional Entropy-Based External Cluster Evaluation Measure*, in: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), 2007, 410–420.

[19] P. J. Rousseeuw: *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, Journal of computational and applied mathematics 20 (1987), pp. 53–65.

[20] Tausczik, Yla R and Pennebaker, James W: *The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods*, Journal of Language and Social Psychology 29.1 (2010), 24–54.

[21] R. Verma, N. Shashidhar, N. Hossain: *Detecting Phishing Emails the Natural Language Way*, in: Computer Security – ESORICS 2012, Springer Berlin Heidelberg, 2012, 824–841, isbn: 978-3-642-33167-1.