International Journal of Technology 13(2) 345-355 (2022)
Received August 2020 / Revised February 2021 / Accepted August 2021

International Journal of Technology

http://ijtech.eng.ui.ac.id

Authorship Obfuscation System Development based on Long Short-term Memory Algorithm

Hendrik Maulana[1*], Riri Fitri Sari[1]

[1]*Department of Electrical Engineering, Universitas Indonesia, PO Box 16424, Indonesia*

Abstract. Stylometry is an authorship analysis technique that uses statistics. Through stylometry, the authorship identity of a document can be analyzed with high accuracy. This poses a threat to the privacy of the author. Meanwhile, there is a stylometry method, namely the elimination of authorship identity, which can provide privacy protection for writers. This study uses the authorship method to eliminate the method applied to the Federalist Paper corpus. Federalist Paper is a well-known corpus that has been extensively studied, especially in authorship identification methods, considering that there are 12 disputed texts in the corpus. One identification method is the use of the support vector machine (SVM) algorithm. Through this algorithm, the author's identity of disputed text can be obtained with 86% accuracy. The authorship identity elimination method can change the writing style while maintaining its meaning. Long-short-term memory (LSTM) is a deep learning-based algorithm that can predict words well. Through a model formed from the LSTM algorithm, the writing style of the disputed documents in the Federalist Paper can be changed. As a result, 4 out of 12 disputed documents can be changed from one author identity to another identity. The similarity level of the changed documents ranges from 40% to 57%, which indicates the meaning preservation from original documents. Our experimental results conclude that the proposed method can eliminate authorship identity well.

Keywords: Authorship; Long-short-term memory (LSTM); Obfuscation

1. Introduction

Stylometry is a science that analyzes authorship style using statistics. Most research in the field of stylometry refers to Mosteller and Wallace's research on Federalist Papers in 1963. With the beginning of computer-based stylometry, the corpus of Federalist Papers gained popularity. Stylometry is classified into several working subsections, namely authorship identification, authorship verification, authorship profile, stylochronometry, and authorship elimination, with the majority of studies in the first three classes. In contrast, authorship deletion is generally used for concealing identity when an author does not want their identity to be revealed publicly. Authorship identification methods have progressed rapidly, currently achieving an accuracy of up to 90% (Iqbal et al., 2020). This rapid development has raised serious threats to privacy for certain professionals, such as journalists and activists.

*Corresponding author's email: riri@eng.ui.ac.id; hendrik.maulana@ui.ac.id, Tel.: +62-85726191091
doi: 10.14716/ijtech.v13i2.4257

McDonald et al. (2012) proposed a method of authorship style transformation that can be entered manually by the user to make their writing anonymous. Due to the limitations of manual techniques, Almishari et al. (2013) proposed a machine translation method to automate authorship style changes. Other previously proposed methods have studied the use of synonyms, sentence separation, sentence combinations, and paraphrases, but all studies in this field still have misspellings in their text results. In addition, no single study provides randomness settings for users. This paper proposes a method that has better text results and gives text randomness control for the user.

Research in the field of authorship identity elimination requires a standardized dataset. The dataset must be well known, have anonymous documents, and should have been identified by the number of previous authorship attribution studies. One of the datasets that matches this criterion is the Federalist Paper. The corpus contains 85 documents, including 51 documents written by Hamilton, 14 documents written by Madison, 5 documents written by Jay, 3 documents are a collaboration between Hamilton and Madison, the remaining 12 documents are doubtful of their authorship between Hamilton or Madison. Juola (2020) identified no fewer than 19 studies conducted on this corpus. One of them was conducted by Savoy (2013), who applied several algorithms and concluded that only the Naive Bayes and SVM algorithms have relevant results. SVM is one of the encouraging classification techniques in the field of machine learning (Abdillah, et al., 2016).

Rahguoy et al. (2018) conducted sentence separations, combining sentences and replacing words through WordNet. WordNet is a lexical database used to extract features from sentences (Santosh, et al., 2015). This method can reduce the confidence level in the authorship identification process by 20%, but not all sentences produced can be arranged properly. Other attributes examined by Karadzhov et al. (2017) are the ratio of word types, stopword ratio, ratio of large capitalized words, and the ratio of part of speech. In contrast, Bakhteev and Khazov (2017) changed the sentence level by paraphrasing and modifying the content using the LSTM algorithm through the encoder–decoder technique. As a result, they obtained a fairly high sentence change rate, but there were still some spelling mistakes.

Two requirements need to be fulfilled for the authorship elimination method: Confirmation of authorship identity modification to be proven by SVM classification and confirmation of meaning preservation. The semantic similarity method can show the meaning preservation, because the greater the value of semantic similarity, the greater is the similarity of meaning between two documents (Sitikhu et al., 2019). There are techniques that can be used to calculate the semantic similarity of documents, such as the Jaccard coefficient, dice coefficient, and cosine similarity. Afzali and Kumar (2017) found that cosine similarity is the best performance evaluation technique for calculating semantic similarity. Deleting authorship identity in this research can be used to test the performance of the authorship identification method if the document has been changed. The long-short term memory (LSTM) algorithm based on a neural network is used as a rearrangement of disputed documents because of its suitable performance in natural language generation (Lippi et al., 2019).

2. Methods

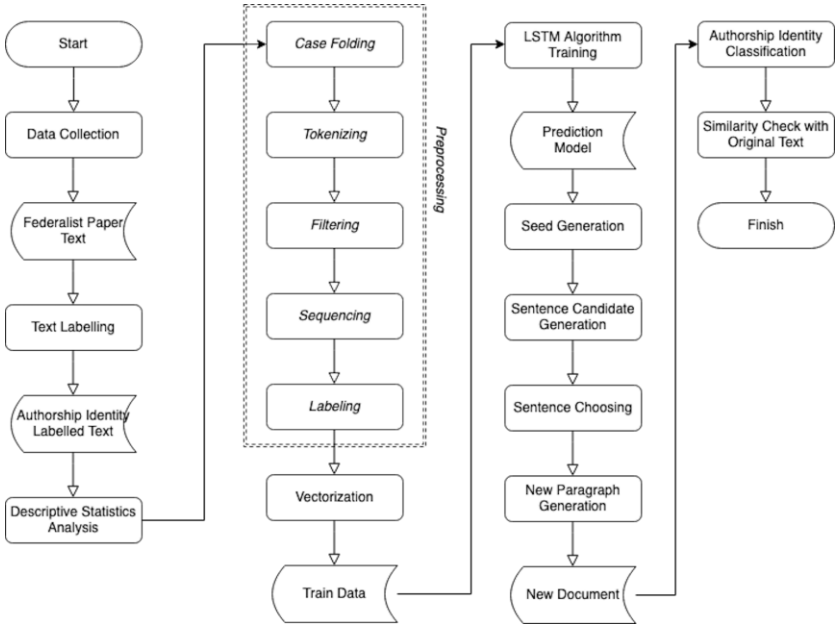

Figure 1 Authorship elimination system flowchart

2.1. Sample

This study uses the purposive sampling technique to determine samples. The samples are obtained through the inclusion criteria determined by the researcher. The inclusion criteria are documents whose authorship identity is debated, whether written by Madison or Hamilton, while the exclusion criteria are data that are not used in the study. The exclusion criteria are documents written by John Jay.

2.2. Analysis Method

This research uses descriptive statistical analysis methods to describe the writing style in Federalist Paper documents. The descriptive statistical analysis method will show the characteristics of writing style quantifications, such as frequently appearing words, sentence length, word distribution, and vocabulary similarity. The first requirement for eliminating authorship identities is to be able to show a change in writing style. To analyze this change, a classification algorithm is required to show the differences in the classification results before and after modifying the writing style. A support vector machine (SVM) is a machine learning algorithm that can perform text classification. In this study, SVM is used because of its accuracy in overcoming other classification algorithms, such as decision trees, OPR, and OLR (Grishunin, et al., 2020).

$$\frac{1}{2}||w||^2 + C\Sigma_{i=1}^{n}\xi_i \quad (1)$$

The word vector is used as input for the SVM algorithm. Due to the large number of documents written by Hamilton, the text data are normalized by adding a sample of Madison's text so that the number is the same as Hamilton's. Then, the vector is separated into training data and test data. After training on the SVM model, the value of the ROC curve is close to 1.0. This indicates that the trained SVM model can classify document authorship identities well.

The second requirement for eliminating authorship identities is being able to maintain document meaning. The cosine similarity algorithm can show the meaning similarity of two documents and thus can be used to determine the meaning preservation of authorship elimination results. Moreover, it performs better than the Jaccard coefficient and the dice coefficient.

$$cos\ \theta\ = \frac{A.B}{||A||||B||} \quad (2)$$

To modify the writing style, an approach that can change the sentence automatically is required. LSTM is one of the best-performing algorithms in language modeling. It can be used to modify the writing style in the elimination of authorship identity. The sentence vector is used as input to LSTM to preserve sentence meaning. Therefore, changes in writing style do not change the original meaning of the document; modifying the writing style is combined with a synonym. Bakhteev and Khazov (2017) used the LSTM model with an encoder–decoder approach. This paper proposes a different method that utilizes randomness settings in word order.

2.3. Authorship Elimination System Design

This research is divided into three phases, as shown in Figure 1. The first phase is processing data in the form of text obtained from Federalist Paper documents. Descriptive statistical analysis methods are applied to the data to obtain words that often appear and that can provide a distinct authorship identity. The labeled data are then preprocessed, which includes tokenization and filtering. According to a traditional analysis, Madison wrote the disputed documents.

The second phase of this research involves modification of the writing style. The training data are documents written by Madison and Hamilton. Text preprocessing is performed through tokenization and filtering. After that, tokens from the text are grouped sequentially. Then, the sequence data in the form of word tokens are converted to numeric values through vectorization. From the numerical sequence data, they are then converted into a matrix and entered into the LSTM algorithm. This algorithm predicts the next word from the formed sequence. To improve the quality of the text produced, the authors utilize the Doc2Vec model, which can see the context per sentence so that the meaning of each sentence can be maintained. The obtained model is then applied to the test data, namely the sequences of the disputed document.

The third phase involves testing and interpreting the results. The prediction text produced by the LSTM algorithm-based model is tested through authorship classification based on the SVM algorithm. Then, the transformed text is measured by cosine similarity to show the preservation of meaning from original documents. From this process, we can determine the change in authorship identity and the similarity of the documents to the original documents to fulfill the requirements of the authorship elimination method.

3. Results and Discussion

3.1. Descriptive Analysis

In this study, descriptive analysis is conducted to obtain a general picture of the information in the Federalist Paper document. Depending on the author, the text of the Federalist Paper is divided into four groups, namely documents written by Hamilton, Madison, and Jay, as well as those debated by the author. Hamilton wrote most of the documents on Federalist Papers, as shown in Figure 2.

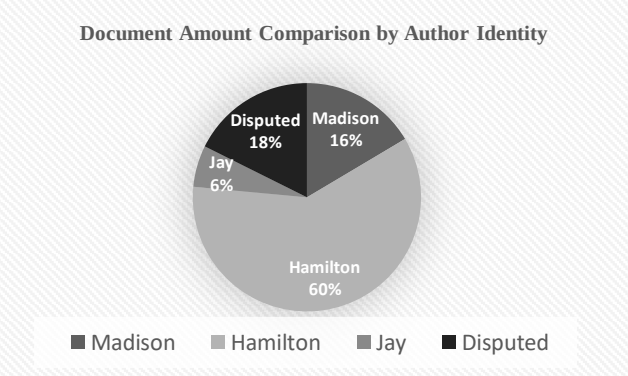

Figure 2 Comparison of the number of documents by author identity

In this analysis, we also include a comparison of sentence and word length in the Federalist Paper. The longest document is number 83 with 163 sentences and 6314 words, while the shortest document is number 13 with 28 sentences and 1048 words.

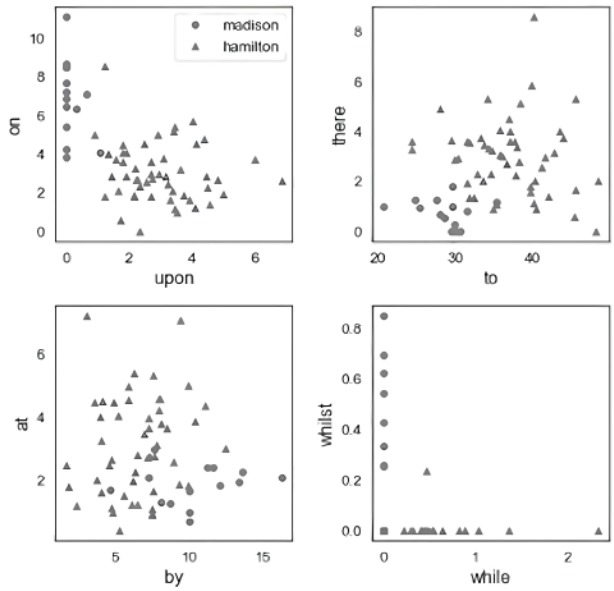

Figure 3 Word choice from Hamilton and Madison documents

The range of word length in one sentence is approximately 20–50 words. Sentences written by Madison have a mean value of 37.7 words, and those written by Hamilton have a mean value of approximately 37.6 words. In the disputed documents, the word mean is 33.8 words per sentence. To determine the word choice of Madison and Hamilton, we compare them using frequent word comparison, as shown in Figure 3. The result shows that Madison uses words such as on, to, by, and whilst, whereas Hamilton prefers upon, there, at, and while.

Khomytska et al. (2020) analyzed the use of chi-square methods to identify the authorship of documents. The smaller the chi-square value, the more similar are the two documents. When two documents are similar, the author should be the same person. From the chi-square value, Madison tends to write the disputed document.

3.2. Text Preprocessing

There are several processes involved in text processing. The first is the case folding, which transforms the capital word into a non-capital word. The second is the process of extracting words from documents that are called tokenizing. This results in a vocabulary of unique words. The third is the process of removing the noise from documents, such as header, footer, space character, and newline. This process is called filtering. The fourth is

the sequencing process, which parses the document into a chunk of word sequences comprising 15 words per sequence.

3.3. Text Representation

The LSTM algorithm used in this research has input as the number, so the text must be converted into numeric form. One of the methods that can be used to achieve this is one-hot encoding. It can be used to transform a text into a sparse matrix consisting of numbers 0 and 1. A sparse matrix is used in the word prediction model. Then, the text is chunked into sequences of 15 words.

Unlike one-hot encoding, Doc2Vec operates at the sentence level, so it will be used in the sentence prediction model. The PV-DBOW model will be used to vectorize sentences because of computation efficiency. The PV-DBOW model is trained by the corpus and results in a model that contains vectors of sentences. Then, the vector is inferred from the PV-DBOW model to obtain the matrix for the LSTM input. In the sentence prediction process, we use five sentences per sequence; this sequence is inferred from the PV-DBOW model to transform them into numeric form.

3.4. Model Architecture

Two models are developed in this study: the word prediction model and sentence prediction model. This study uses three LSTM layers because, according to Merity et al. (2018), it is the effective layer number of an LSTM language model.

An LSTM network comprises three layers: input, hidden, and output. To build the model, we use the Keras library in Python. The input layer is a 3D matrix that comprises samples (batch_size), timesteps, and features. The word prediction model input layer comprises a matrix of dimensions of 20 x 15 x 160845, which results from one-hot encoding vectorization. The sentence prediction model input layer comprises a matrix with dimensions of 20 x 5 x 500, which results from PV-DBOW vectorization. Dropout regularization is also applied in the model to avoid overfitting. The activation function in the word prediction model is added by the temperature parameter to variate the randomness of word prediction.

3.5. Model Training

In the model training process, a search is performed to use the appropriate parameters in the model so that the resulting validation loss can be as minimal as possible. The training process is monitored and stopped in the middle if the value of validation loss is not reduced during five epochs. This is done to avoid overfitting that is too significant. All models are trained with a maximum epoch value of 100. At the end of parameter tuning, several parameters yield the optimum (minimum) validation loss, as shown in Table 1.

Table 1 Optimum parameter value

Model	Neuron	Dropout	Epoch	Batch	Learning Rate	Optimization
Word Prediction	512	0,5	100	20	0.001	Adam
Sentence Prediction	512	0.3	100	20	0.1	SGD

Based on the parameters obtained previously, the training process is presented in the form of a comparison graph of loss values. A good model is obtained when the training loss lines are close to the validation loss lines.

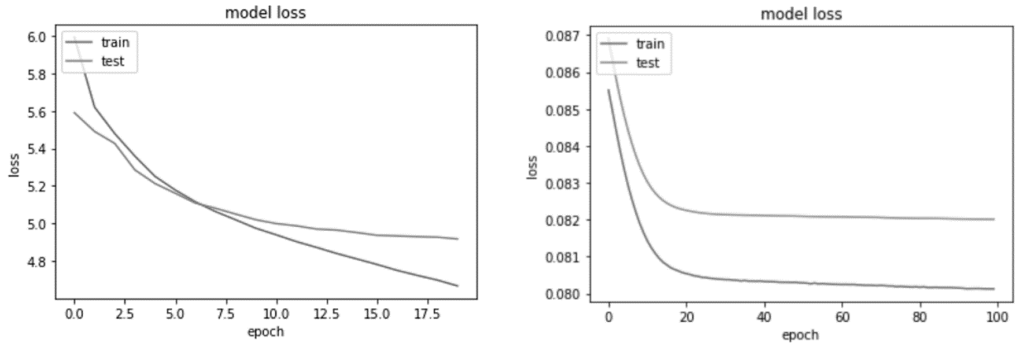

Figure 4 Training model graph on word prediction model (left) and sentence prediction model (right)

The training process can produce a suitable model. Figure **4** shows that the word prediction model does not experience overfitting or underfitting because the lines between the training and testing models are close.

3.6. Modifying the Writing Style

This study uses a different approach to that adopted by Bakhteev and Khazov (2017), who removed authorship identity based on the LSTM algorithm in the encoder–decoder model. This study uses sentence prediction based on the LSTM algorithm by utilizing randomness settings in word order. The randomness setting was achieved using temperature parameters in the LSTM network architecture. Every 5th order of sentences is replaced with the new sentences generated from the LSTM model, and other sentences are modified using a synonym.

3.6.1. Changing words with their synonyms

To maintain the meaning of the original text, not all sentences are rearranged using model predictions. Only certain sentences are rearranged based on the predictions. The remaining sentences are changed using synonyms.

Moesteller and Wallace (1963) used a list of function words, amounting to 70 and 165 words, to distinguish between Madison's and Hamilton's writings. Referring to the word list, a synonym search is performed on the Thesaurus website. To add randomness, the synonym of a word is chosen randomly from the synonym search results on the website.

3.6.2. Arrangement of new sentence candidates

If a sentence is replaced with a new sentence that is predicted, then the prediction model is used. At this stage, several new sentence candidates are arranged as substitutes for the original sentence through the word_prediction model. This process involves the following three steps.

The first is seed arrangement. The text from the dataset needs to be preprocessed to become a seed. In this step, padding is performed if the text taken from the dataset is less than 15 words, and then the tokenization process begins.

The second is the prediction of the next word. The length of the seed used is 15 words, and then the model predicts the 16th word. This step uses the temperature parameter as a randomness variable. If the value of temperature = 1.0, the chance value of the selected word is the same as that of the word that should be next in the sequence. If the temperature value exceeds 1.0, more words can be chosen as the next word. Meanwhile, if the temperature value is less than 1.0, then the word with a small chance value will be ignored, so that the choice of words decreases.

The third is sentence arrangement. The prediction results in words. Then, if a dot, question marks, or exclamation points result from the prediction, the process is stopped and then the predicted words is counted as one sentence.

3.6.3. Selection of best sentence candidates

From the candidate sentences that have been generated, it is necessary to choose the best sentence. Each sentence candidate vectorization value is compared to its similarity with the predicted vector Doc2Vec. The best sentence is chosen based on the highest similarity value.

3.7. Authorship Identification Test

In this study, the SVM algorithm is used to train authorship to identify authorship from debated documents. After training on the model, the results obtained from the model's F1 value are 85% in the Hamilton text and 86% in the Madison text. Furthermore, the model is used to predict authorship identity in the document being debated, and the results of the model indicate that all documents being debated were written by Madison.

Table 2 Confusion matrix of SVM model

	Precision	recall	f1-score	support
0	0.95	0.77	0.85	137
1	0.79	0.95	0.86	124
accuracy			0.86	261
macro avg	0.07	0.86	0.86	261
weighted avg	0.87	0.86	0.86	261

The resulting model is used to identify authorship in documents whose writing style has been changed. The purpose of modifying the writing style is to determine the change in the classification results so that the document originally classified by Madison will be changed to be written by Hamilton. To see the effect of these changes, three parameters are used: the order of sentences replaced, the number of words that are synonymous, and the temperature value.

Table 3 Comparison of classification results after modifying the writing style

Text	Early Classification	End Classification
49	Madison	Madison
50	Madison	Madison
51	Madison	Madison
52	Madison	Madison
53	Madison	Madison
54	Madison	Hamilton
55	Madison	Madison
56	Madison	Hamilton
57	Madison	Madison
58	Madison	Madison
62	Madison	Hamilton
63	Madison	Hamilton

Table **3** shows that after modifying the writing style, the authorship identification of the document number 54, 56, 62, and 63 is changed from the earlier classification. It shows that the accuracy of the SVM model is reduced by approximately 19%.

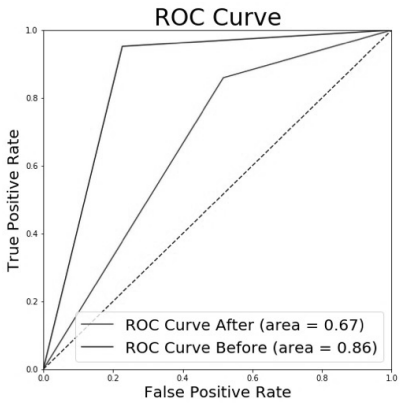

Figure 5 ROC curve comparison after modifying the writing style

This result is better than that obtained by Kacmarcik and Gamon (2006) model, as shown in the table below.

Table 4 Comparison of SVM Accuracy Evaluation Results

Methods	SVM Accuracy
LSTM	67%
Feature Elimination	74%

3.8. Similarity Test

To measure the value of similarity in meaning to the text, the cosine similarity method is used. This method requires vector data input. Therefore, text data must be converted into vectors using the TF-IDF method. The similarity value is obtained from the vector comparison of the changed document to the original document.

Table 5 Similarity comparison between changed and original documents

Document	Similarity Score
49	0,4038
50	0,4998
51	0,5389
52	0,5362
53	0,5779
54	0,4635
55	0,4974
56	0,5172
57	0,5057
58	0,5223
62	0,3878
63	0,4155

Table 5 shows that the similarity of the changed and original documents varies from 39% to 57%. This means that the changed documents still preserve the semantics of the original documents.

4. Conclusions

The analysis conducted in this study indicates that the Federalist Paper corpus is an unbalanced dataset because most articles (60%) were written by Hamilton. Therefore, normalization of text data is required when they are classified. Chi-square and cosine similarity methods show the tendency of the author's to identify a text. If two texts are of high similarity value, then the possibility of the authors of both texts being the same person

is high. Through the SVM algorithm, it is known that the text writer debated in the Federalist Paper can be identified with an accuracy of 86%. Then, using the LSTM algorithm, the level of accuracy can be reduced by 19%. Document changes resulting from the elimination of authorship identity have a similarity level of 39%–57% of the original document, which illustrates that the meaning of the document experiences insignificant changes. These results indicate that authorship identity elimination using the LSTM algorithm achieves suitable performance. For future research, a grid search in parameter tuning can be used to obtain better LSTM parameters. Another method of modifying the writing style can be combined with LSTM text generation, such as separating and combining sentences.

Acknowledgements

This work is supported by Universitas Indonesia under the Q1Q2 Grant Number NKB-0321/UN2.R3.1/HKP.05.00/2019.

References

Afzali, M., Kumar, S., 2017. Comparative Analysis of Various Similarity Measures for Finding Similarity of Two Documents. *International Journal of Database Theory and Application*, Volume 10(2), pp. 23–20

Abdillah, A., Suwarno. 2016. Diagnosis of Diabetes Using Support Vector Machines with Radial Basis Function Kernels. *International Journal of Technology*, Volume 7(5), pp. 849–858

Almishari, M., Gasti, P., Tsudik, G., Oguz, E. (2013). Privacy-Preserving Matching of Community-Contributed Content. *In:* Crampton, J., Jajodia, S., Mayes, K. (eds) Computer Security – ESORICS 2013. ESORICS 2013. Lecture Notes in Computer Science, vol 8134. Springer, Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-40203-6_25

Bakhteev, O., Khazov, A., 2017. Author Masking using Sequence-to-Sequence Models. *In:* 2017 CEUR Workshop Proceedings, http://ceur-ws.org/Vol-1866/paper_68.pdf.

Grishunin, S., Suloeva, S., Egorova, A., Burova, E., 2020. Comparison of Empirical Methods for the Reproduction of Global Manufacturing Companies Credit Ratings. *International Journal of Technology*, Volume 11(6), pp. 1223–1232

Iqbal, F., Debbabi, M., Fung, B., 2020. Authorship Analysis Approaches. *Machine Learning for Authorship Attribution and Cyber Forensics*, pp. 45–56

Juola, P., 2020. Authorship Studies and the Dark Side of Social Media Analytics. *Journal of Computer Science*, Volume 26(1), pp. 156–170

Kacmarcik, G., Gamon, M., 2006. Obfuscating Document Stylometry to Preserve Author Anonymity. *In:* 21st International Conference on Computational Linguistics pp. 444–451

Karadzhov, G., Mihaylova, T., Kiprov, Y., Georgiev, G., 2017. The Case for Being Average: A Mediocrity Approach to Style Masking and Author Obfuscation. *In:* International Conference of the Cross-Language Evaluation Forum for European Languages, https://doi.org/10.48550/arXiv.1707.03736.

Khomytska, I., Teslyuk, V., Kryvinska, N., Bazylevych, I. 2020. Software-Based Approach towards Automated Authorship Acknowledgement—Chi-Square Test on One Consonant Group. *Electronics*, Volume 9(7), pp. 1–11

Lippi, A., Montemurro M., Esposti, M., Cristadoro, G., 2019. Natural Language Statistical Features of LSTM-Generated Texts. *IEEE Transactions on Neural Networks and Learning Systems*, Volume 30

Moesteller, F., Wallace, D., 1963. Inference in an Authorship Problem. *Journal of the American Statistical Association*, Volume 58(302), pp. 275–309

McDonald, A., Afroz, S., Caliskan, A., Stole, A., 2012. Use Fewer Instances of the Letter "i": Toward Writing Style Anonymization. *In:* International Symposium on Privacy Enhancing Technologies, https://doi.org/10.1007/978-3-642-31680-7_16.

Merity, S., Keskar, N., Socher, R., 2018. Regularizing and Optimizing LSTM Language Models. *In:* International Conference on Learning Representations

Rahguoy, M., Giglou, H., Rahguoy, T., Zaeynali, H., 2018. Author Masking Directed by Author's Style. *In:* 2018 CEUR Workshop Proceedings, https://pan.webis.de/downloads/publications/papers/rahgouy_2018.pdf.

Santosh, D., Vardhan, B., 2015. Obtaining Feature and Sentiment-Based Linked Instance RDF Data from Unstructured Reviews using Ontology-Based Machine Learning. *International Journal of Technology*, Volume 2, pp. 198–206

Savoy, J., 2013. The Federalist Papers Revisited: A Collaborative Attribution Scheme. *In:* Proceedings of The American Society for Information Science and Technology, Volume 50(1), pp. 1–8

Sitikhu, P., Pahi, K., Thapa, P., Shakya, S., 2019. A Comparison of Semantic Similarity Methods for Maximum Human Interperability. *In:* IEEE International Conference on Artificial Intelligence for Transforming Business and Society, https://doi.org/10.48550/arXiv.1910.09129