Forensic Authorship Classification by Paragraph Vectors of Speech Transcriptions

Dávid Sztahó[1], András Beke[1], György Szaszák[1], Attila Fejes[2]

[1] Budapest University of Technology and Economics,
Magyar tudósok körútja 2., 1117 Budapest, Hungary
sztaho.david@vik.bme.hu, beke.andras@lsa.tmit.bme.hu,
szaszak.gyorgy@vik.bme.hu
[2] University of Public Sevice Doctoral School of Law Enforcement,
Ludovika tér 2., 1083 Budapest, Hungary
fejes.attila@nbsz.gov.hu

Abstract: In forensic comparison, document classification techniques are used mainly for authorship classification and author profiling. In the present study, we aim to introduce paragraph vector modelling (by Doc2Vec) into the likelihood-ratio framework paradigm of forensic evidence comparison. Transcriptions of spontaneous speech recording are used as input to paragraph vector extraction model training. Logistic regression models are trained based on cosine distances of paragraph vector pairs to predict the same and different author origin probability. Results are evaluated according to different speaking styles (transcriptions of speech tasks available in the dataset). C_{llr} and equal error rate values (lowest ones are 0.47 and 0.11, respectively) show that the method can be useful as a feature for forensic authorship comparison and may extend the voice comparison methods for speaker verification.

1 Introduction

In forensic comparison practice, a widely spreading automatic evaluation method is getting more and more accepted. The method is called likelihood ratio (LR) framework (Morrison, 2011; Saks and Koehler, 2005). It emerges in numerous topics, most widely known in DNA identification. The key point of the paradigm is to resolve the question: how characteristics are given measures to an individual or to a population. In the course of actual application, it considers two hypotheses: "What is the probability that the sample in question comes from the suspect?" And the so-called counter-hypothesis: "What is the probability that the sample in question comes from another person randomly selected from a given population?". Based on these, the probability of the evidence can be written:

$$LR = \frac{p(E|H_{so})}{p(E|H_{do})} \quad (1)$$

where LR is the likelihood-ratio, E is the evidence, H_{so} is the hypothesis of same-origin subjects, H_{do} is the hypothesis of different-origin subjects.

In forensics, natural language processing (NLP) methods target mainly authorship classification (Khonji et al., 2021) and author profiling in which features of a person such as gender, age, and cultural characteristics are classified. The basic method for both methods is to extract single features (such as vocabulary size, word frequency) from a document and apply a machine learning method (such as support vector machines) for classification (Adame-Arcia et al., 2017; Estival et al., 2007; Hsieh et al., 2018). As deep learning methods emerged in NLP, two main document classification methods became more widespread: Doc2Vec and BERT language modelling. Doc2Vec (Le and Mikolov, 2014) is an extension of Word2Vec (Mikolov et al., 2013) and is used basically for document classification and document similarity scoring. It was also applied for author profiling (Markov et al., 2016) detection certain speaker characteristics (age, gender) but not the speaker ids directly. This method replaced hand-crafted feature extraction by automatic modelling trained on a large corpus. In this method the Word2Vec method is extended by a vector called document vector (Fig. 1) that is trained along with the vector representation of words (Word2Vec, (Mikolov et al., 2013)). This document vector will contain an accumulated information about the paragraphs for a single document.

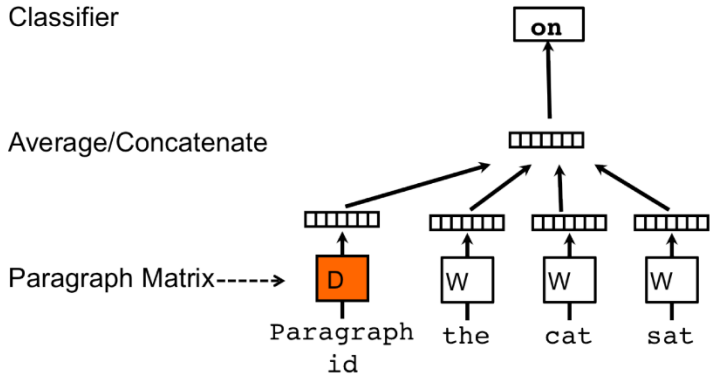

Fig. 1. Deep learning architecture for Doc2Vec (Le and Mikolov, 2014)

The present work aims to insert document classification (by Doc2Vec) into the LR framework of forensic authorship verification (as a text-based speaker recognition) based on transcriptions of spontaneous speech. To our knowledge, this type of LR authorship verification based on paragraph vectors has not been done before. The transcriptions are created based on samples of the ForVOICE project (Beke et al., 2021) containing free dialogues and monologues. Logistic regression models were created for same and different speaker probabilities (Eq. 1). The results are evaluated by C_{llr} end equal error rate scores and plotted on tippet plots common in forensic evidence comparison. The method can be used to combine text-based features and voice-based features in order to improve overall speaker verification results.

2 Methods

2.1 Transcriptions

The trained models were evaluated using the ForVOICE dataset (Beke et al., 2021). It contains spontaneous speech of three different speaking styles: (1) free dialogue (~10 minutes), (2) guided dialogue (~8 minutes) and (3) monologues (~3 minutes). 80 speakers were recorded twice (with 2 weeks interval apart) and transcriptions of all speech samples were created manually. 60 speakers were randomly selected for model training and the remaining 20 speakers were used for evaluation.

Beside the 60 speakers of the ForVOICE dataset, transcriptions of spontaneous speech were used from the BEA (Gósy et al., 2012) and the HuComTech (Szekrényes, 2014) datasets. The total number of words and paragraphs used is shown in Table 1.

Training data were created in two ways from transcriptions: (1) splitting every sample transcription to word lengths of 200 and using 100 words overlap (multiple fixed length paragraphs for a speech sample) and (2) using every sample transcription as a single training paragraph (a single variable length paragraph for each speech sample). In the former case, multiple paragraphs are available for a given recording. For example, the original transcription of the first monologue recording of a speaker is split into multiple overlapping parts with 200 word lengths and is used in the experiments. In the latter case, the original total transcription is used. Same splitting is done for the BEA and the HuComTech transcriptions.

Table 1: Number or words and paragraphs in corpora used

Dataset	#speakers	#words (text lengths)	#paragraphs with splitting	#paragraphs without splitting
Train: ForVOICE	60	163874	1813	360
Train: ForVOICE + BEA + HuComTech	182	352238	3865	773
Test: ForVOICE	20	127208	1407	120

2.2 Doc2Vec modelling

Doc2Vec (Le and Mikolov, 2014) approach is implemented by the Gensim Python package (version 4.1.2). In this method the Word2Vec method is extended by a vector called document vector that is trained along with the vector representation of words (Word2Vec, (Mikolov et al., 2013)). This document vector will contain an accumulated information about the paragraphs for a single document. Vocabulary for the Doc2Vec method is built from the training dataset. Due to the limited training corpus size available from spontaneous speech, this results that 37% and 27% of the test vocabulary

entries were not covered by train vocabulary entries in case of ForVOICE and ForVOICE + BEA + HuComTech cases, respectively.

Doc2Vec models were built using different paragraph vector length (20, 100, 200) and training epoch number (40, 100, 200) using speaker ids. The maximum distance between the current and predicted word within a sentence was set to 12 based on (Markov et al., 2016). After the model is created, paragraph vectors extracted on the evaluation speaker set were used for same speaker and different speaker origin modelling by logistic regression (LR, implemented by Python sklearn package). LR models are built using cosine distances of extracted paragraph vectors as input using the target variable if a vector pair is of same or different speaker origin. The output of the LR model is the probability of the same speaker decision. This enables the calculation of Eq. 1. Fig. 2 shows a sample of a trained LR model. Distribution of same and different origin vector pairs are shown in yellow and blue, respectively.

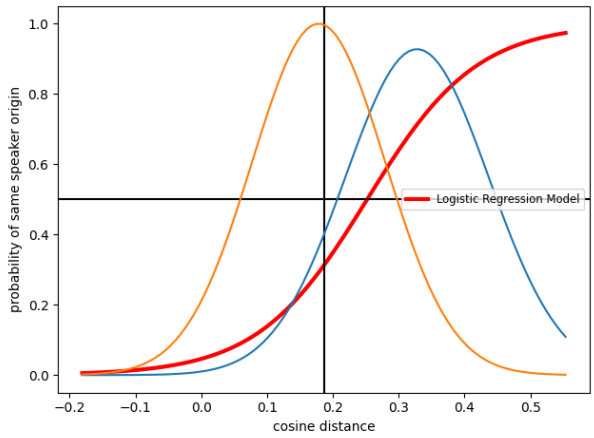

Fig. 2. A trained LR model sample. Yellow and blue lines show the distributions of paragraph vector pairs of same and different author origin.

2.3 Evaluation

Models (Doc2Vec and LR) were created on the training dataset (training speaker set of ForVOICE with and without augmentation by BEA and HuComTech) and evaluated using the test speaker set of ForVOICE. Evaluation metrics are equal error rate (EER) of author verification (EER is the level where false acceptance rate and false rejection rate are equal, commonly used in biometric security systems) and log-likelihood-ratio cost (C_{llr},, Eq. 2) (Van Leeuwen and Brümmer, 2007), defined as

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{so}}\sum_{i=1}^{N_{so}}\log_2\left(1+\frac{1}{LR_{so_i}}\right)+\frac{1}{N_{do}}\sum_{j=1}^{N_{do}}\log_2\left(1+LR_{do_j}\right)\right)$$

where N_{so} and N_{do} are the number of same-origin and different-origin comparisons and LR_{so} and LR_{do} are the likelihood ratios derived from same-origin and different-origin comparisons. C_{llr} is a function measuring the balance of LR scores of same-origin and

different-origin comparisons measured using all possible same-origin and differenc-origin vector pair combinations. Ideal same-origin and different-origin comparisons have logLR>0 and logLR<0, respectively. Incorrect (not as ideal as the mentioned inequalities) produce a higher C_{llr}. The better the performance of a forensic comparison system, the more correct *LR* values are produced, the lower C_{llr} is achieved, supplying the evidence magnitude.

3 Results

Doc2Vec and LR models were created for various speech task types to evaluate if different speaking styles (domains) affect authorship verification performance. Four cases were considered: using texts from all speech tasks altogether and from each single tasks individually. Doc2Vec models were trained in two variations: texts of the total ForVOICE corpus (of speakers selected for training) with and without augmentation by BEA and HuComTech corpora. Logistic regression models were always trained on texts of the given speech styles, while Doc2Vec models were created without speech task filters. Tables 2 and 3 contain C_{llr} and EER values (without and with splitting paragraphs into word lengths of 200, respectively) calculated on the test speaker set of the ForVOICE corpus for all cases, paragraph vector lengths and training epoch numbers. Best cases for each speech tasks and training corpora based on the C_{llr} and EER values are highlighted. EER and C_{llr} values should not necessarily be perfectly correlated, so there are occuencies where a case with higher C_{llr} achieves lower EER and vice versa. A higher C_{llr} means a shift in the threshold of same-different origin speaker decision from the optimal 0 value. It may indicate a worse generalization level.

It is clear from the tables that LR models trained on single speaking styles achieve better performance than using all tasks at once. There seems to be no significant difference between splitted and non-splitted paragraphs and also training corpora augmentation shows no real difference. The best C_{llr} values (and EERs) for speaking styles are: ‘all’ - 0.87 (0.35); 1 - 0.64 (0.2); 2 - 0.47 (0.11); 3 - 0.65 (0.15). Tippet plots, showing the proportion of correctly identified same and different author origin (commonly used plots in forensic comparison) of the aforementioned results are shown in Fig. 3. Blue and yellow lines show the proportion of correctly identified same and different author origin vector pairs as function of logLR score thresholds. While the blue line measures the proportion of vector pairs belonging to different origin below the given logLR threshold, the yellow line depicts the proportion of the vector pairs belonging to same origin above the given logLR threshold. EER is measured at the crossing of the two lines.

Table 2: Result without splitting paragraphs

training corpus	task	vector	epoch 40		epoch 100		epoch 200	
			eer	C_{llr}	eer	C_{llr}	eer	C_{llr}
ForVOICE	all	20	0,4	0,94	0,44	0,97	0,44	0,97
		100	**0,35**	**0,87**	0,36	0,94	0,41	1,03
		200	**0,35**	**0,87**	**0,34**	**0,93**	0,39	1,09
	1	20	0,3	1,1	0,25	0,93	0,31	1,01
		100	0,3	0,87	0,25	0,86	0,3	0,94
		200	0,2	0,79	**0,2**	**0,64**	0,25	1,04
	2	20	0,25	0,79	0,2	0,71	0,17	0,59
		100	0,19	0,72	**0,11**	**0,6**	0,15	0,62
		200	0,2	0,8	**0,1**	**0,6**	0,14	0,86
	3	20	0,25	0,84	0,31	0,99	0,3	0,94
		100	0,2	0,77	0,2	0,78	0,26	0,89
		200	**0,15**	**0,69**	0,2	0,74	0,25	1,23
ForVOICE + BEA + HuComTech	all	20	0,41	0,95	0,44	0,98	0,45	0,98
		100	0,34	0,89	0,39	0,94	0,39	0,96
		200	**0,33**	**0,88**	0,37	0,95	0,41	1,05
	1	20	0,25	0,78	0,3	0,89	0,3	0,87
		100	**0,2**	**0,78**	0,25	0,83	0,25	0,77
		200	**0,25**	**0,69**	**0,2**	**0.95**	0,25	1,16
	2	20	0,23	0,69	0,3	0,77	0,31	0,76
		100	0,2	0,71	**0,15**	**0,63**	0,2	0,76
		200	0,25	0,7	0,15	0,7	0,15	0,93
	3	20	0,25	0,76	0,22	0,83	0,25	0,9
		100	0,2	0,72	0,2	0,81	0,25	0,74
		200	**0,15**	**0,66**	0,17	1,03	0,23	1,28

Table 3: Results with splitting paragraphs

training corpus	task	vector	epoch 40		epoch 100		epoch 200	
			eer	C_{llr}	eer	C_{llr}	eer	C_{llr}
ForVOICE	all	20	0,42	0,96	0,45	0,98	0,44	0,98
		100	0,35	0,9	0,41	0,96	0,41	0,99
		200	**0,33**	**0,89**	0,38	0,97	0,41	1,03
	1	20	0,3	0,77	0,25	0,82	0,32	0,77
		100	0,29	0,87	0,3	0,75	0,31	0,86
		200	0,25	0,75	**0,2**	**0,68**	0,26	0,88
	2	20	0,31	0,93	0,25	0,66	0,25	0,71
		100	0,15	0,67	**0,11**	**0,47**	0,1	0,55
		200	0,12	0,58	0,15	0,55	0,15	0,62
	3	20	0,26	0,92	0,31	0,92	0,35	1,01
		100	0,24	0,71	0,2	0,68	0,25	0,77
		200	**0,15**	**0,65**	0,15	0,73	0,23	0,89
ForVOICE + BEA + HuComTech	all	20	0,44	0,98	0,46	0,99	0,45	0,99
		100	0,36	0,91	0,4	0,95	0,41	0,97
		200	**0,34**	**0,9**	0,38	0,95	0,41	0,99
	1	20	0,3	0,85	0,25	0,72	0,3	0,84
		100	0,25	0,82	0,25	0,81	**0,27**	**0,7**
		200	**0.21**	**0.79**	0.25	0,83	0,29	0,87
	2	20	0,3	0,76	0,3	0,84	0,37	0,85
		100	0,15	0,65	0,2	0,72	**0,15**	**0,57**
		200	**0,14**	**0,59**	0,2	0,62	**0,15**	**0,57**
	3	20	0,3	0,93	0,3	1,03	0,29	1,02
		100	0,2	0,71	0,2	0,7	0,25	0,81
		200	**0,15**	**0,66**	0,2	0,76	0,2	0,82

4 Discussion and Conclusion

Based on the results, it can be stated that the paragraph vector modelling can be used in a forensic authorship verification framework. However, it is not clear right now as to what vector length to choose. It seems that longer vector lengths perform better. Increasing the number of epochs also increases the C_{llr} indicating an overfitting effect. This can be overcome by using a technique such as early stopping in which a development dataset is used to measure performance during training. Thus, a desired generalization ability can be set.

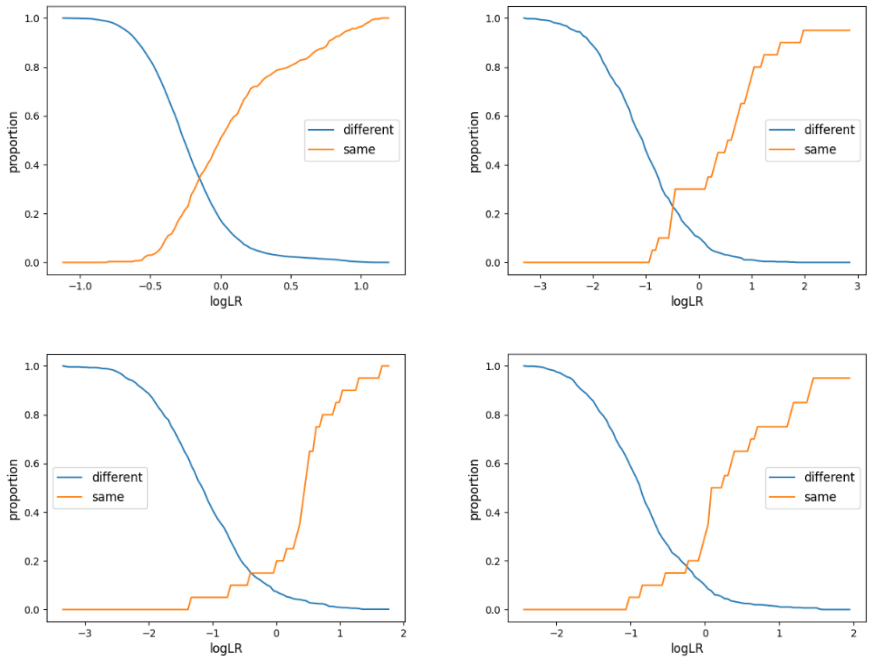

Fig. 3. Tippet plots of best models for speech tasks. From top left to bottom right: 'all', 1, 2 and 3. Blue and yellow lines show the proportion of same and different author origin vector pairs as function of logLR score thresholds, respectively.

It is also clear that domain specific logistic regression models achieve lower C_{llr} and equal error rate values. It may be due to a speaking style mismatch of the speech tasks investigated in the present work.

Comparing current results to related words is hard due to the corpora and target mismatch. In (Kaur et al., 2020), social network posts are identified if they originate from the same user or not. In their work, carefully crafted textual features are used. The current method presented here may serve as an additional feature extending current feature sets, not only in text-based authorship classification/verification but also in forensic voice comparison.

Creating ForVOICE is currently in its final step. Current results and the method will be extended using the final dataset which includes samples of 120 speakers. The final goal is to combine text-based (traditional hand-crafted and deep learning based) features and voice-based features, such as x-vectors, into a final method and evaluate it on a dataset matching forensic needs. Beside Doc2Vec, BERT modelling would also be useful to investigate as a document classification method.

Acknowledgement

The work was funded by project no. FK128615, which has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the FK_18 funding scheme.

References

Adame-Arcia, Y., Castro-Castro, D., Bueno, R.O., Muñoz, R., 2017. Author profiling, instance-based similarity classification. Notebook for PAN at CLEF2.

Beke, A., Szaszák, G., Sztahó, D., 2021. FORvoice 120+: magyar nyelv\Hu utánkövetéses adatbázis kriminalisztikai célú hangösszehasonlításra.

Estival, D., Gaustad, T., Pham, S.B., Radford, W., Hutchinson, B., 2007. Author profiling for English emails, in: Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics. p. 272.

Gósy, M., Gyarmathy, D., Horváth, V., Gráczi, T.E., Beke, A., Neuberger, T., Nikléczy, P., 2012. BEA: Beszélt nyelvi adatbázis.

Hsieh, F., Dias, R., Paraboni, I., 2018. Author profiling from facebook corpora, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Kaur, R., Singh, S., Kumar, H., 2020. TB-CoAuth: Text based continuous authentication for detecting compromised accounts in social networks. Applied Soft Computing 97, 106770.

Khonji, M., Iraqi, Y., Mekouar, L., 2021. Authorship Identification of Electronic Texts. IEEE Access 9, 101124–101146.

Le, Q., Mikolov, T., 2014. Distributed representations of sentences and documents, in: International Conference on Machine Learning. PMLR, pp. 1188–1196.

Markov, I., Gómez-Adorno, H., Posadas-Durán, J.-P., Sidorov, G., Gelbukh, A., 2016. Author profiling with doc2vec neural network-based document embeddings, in: Mexican International Conference on Artificial Intelligence. Springer, pp. 117–131.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems. pp. 3111–3119.

Morrison, G.S., 2011. Measuring the validity and reliability of forensic likelihood-ratio systems. Science & Justice 51, 91–98.

Saks, M.J., Koehler, J.J., 2005. The coming paradigm shift in forensic identification science. Science 309, 892–895. https://doi.org/10.1126/science.1111565

Szekrényes, I., 2014. Annotation and interpretation of prosodic data in the hucomtech corpus for multimodal user interfaces. Journal on Multimodal User Interfaces 8, 143–150.

Van Leeuwen, D.A., Brümmer, N., 2007. An introduction to application-independent evaluation of speaker recognition systems, in: Speaker Classification I. Springer, pp. 330–353. https://doi.org/10.1007/978-3-540-74200-5_19