Student Symposium On Artificial Intelligence

EPIA 2023

22nd EPIA Conference on Artificial Intelligence

September 5, 2023
Faial, Azores, Portugal

Organizers

Raquel Sebastião
Universidade de Aveiro / IEETA, Portugal

Catarina Silva
University of Coimbra / CISUC, Portugal

Nuno Moniz
Lucy Family Institute for Data & Society, University of Notre Dame, USA

Mentors

Alberto Freitas
Faculdade de Medicina, Universidade do Porto, Portugal

Bernardete Ribeiro
Universidade de Coimbra/CISUC, Portugal

Branka Hadji Misheva
Bern University of Applied Sciences, Switzerland

Henrique Lopes Cardoso
Faculdade de Engenharia da Universidade do Porto, Portugal

Hugo Oliveira
Universidade de Coimbra/CISUC, Portugal

João Gama
Faculdade de Economia da Universidade do Porto, Portugal

Joaquim Silva
Faculdade de Ciências e Tecnologia, Universidade Nova, Lisboa, Portugal

Paulo Cortez
Universidade do Minho/ ALGORITMI, Portugal

Paulo Rupino
Universidade de Coimbra/CISUC, Portugal

Petia Giorgieva
Universidade de Aveiro / IEETA, Portugal

Raquel Sebastião
Universidade de Aveiro / IEETA, Portugal

Rita Ribeiro
Faculdade de Ciências da Universidade do Porto, Portugal

Accepted Submissions

AI for Health

Emotional state classification from brain signals using CNNs adapted for fMRI signal properties
Daniel Agostinho

Learning Postoperative Pain through Physiological Signals
Daniela Pais

Electrocardiogram for Biometric Recognition: Collectability, Stability and Application Challenges
Teresa Pereira

AI in Natural Language Processing and Data

Clustering Massive, Noisy, and Unstructured Textual Streams
Cesar Andrade

Data Quality, Data Balance and Data Documentation: a framework
Marco Rondina

Estimating the Density Ratio with a ReLU Induced Tessellation
Michal Lewandowski

A Semantic Search System for the Supremo Tribunal de Justiça
Rui Melo

Responsible AI, Art, and Finance

Addressing Self-Sustainability in Multi-Agent Systems through Combating Terrorism Financing
David Makiya

Data Leakage Detection and Data Denoising using Causal Mechanisms for Recommender Systems
Margarida Costa

Can we hold AGI-enabled Robot Morally Responsible for their Actions?
Mubarak Hussain

Symbolic music generation conditioned on continuous-valued emotions
Serkan Sulun

Exploring GenAI Art Tools By Designing Visual Culture Chatbot Assistant: A Case Study In Higher Education
Yi Yang Liu

AI for Health

EMOTIONAL STATE CLASSIFICATION FROM BRAIN SIGNALS USING CNNS ADAPTED FOR FMRI SPECIFICITIES

Daniel Agostinho$^{1,2,*[0000-0003-1503-2948]}$

[1] Center for Informatics and Systems of University of Coimbra (CISUC), Faculty of Science and Technology, University of Coimbra, Portugal
[2] Coimbra Institute for Biomedical Imaging and Translational Research (CIBIT), ICNAS, Faculty of Medicine, University of Coimbra, Portugal
danielagostinho@dei.uc.pt

Abstract. Accurately classifying emotional states from functional magnetic resonance imaging (fMRI) data presents challenges in the field of brain state classification. Traditional machine learning methods struggle with high-dimensional fMRI data, leading to a growing interest in deep learning (DL) models. In this study, we propose a novel approach to classify emotional arousal levels using fMRI data by adapting the EEGNet architecture, originally designed for electroencephalography (EEG) classification. By leveraging two-dimensional representations of the fMRI time courses, EEGNet models achieved accuracies ranging from 70.49% to 72.54%. Furthermore, our newly developed network, fMRINet, outperforms previous models, reaching an accuracy of 73.76%. These results highlight the potential of DL models to capture complex patterns in fMRI data, even with limited available data. Our findings contribute to the field of brain state classification and provide insights into the classification of different emotional states using fMRI data.

Keywords: Deep Learning, fMRI, Brain Decoding, Emotion Classification

1 Introduction

In recent years, neuroimaging techniques like functional magnetic resonance imaging (fMRI) have revolutionized our understanding of the human brain and its connection to emotions. Emotions, which significantly influence human cognition, behavior, and mental well-being, can become challenging to manage when abnormally experienced due to underlying conditions, potentially leading to specific phobias and mental health problems [13, 14]. Accurate classification of emotional states from fMRI data can therefore become crucial for advancing our understanding of emotional processing and developing novel diagnostic and therapeutic tools. In this context, the support vector machine (SVM)-based multi-voxel pattern analysis (MVPA) has emerged as the most widely used method for brain decoding. However, despite its popularity, the SVM faces challenges in handling high-dimensional raw data and requires expert knowledge

[*] PhD in Informatics Engineering, Department of Informatic Engineering of the University of Coimbra

PhD Advisors:

Miguel Castelo-Branco$^{2[0000-0003-4364-6373]}$ and Marco Simões$^{1,2[0000-0003-3713-2464]}$

1 Center for Informatics and Systems of University of Coimbra (CISUC), Faculty of Science and Technology, University of Coimbra, Portugal
2 Coimbra Institute for Biomedical Imaging and Translational Research (CIBIT), ICNAS, Faculty of Medicine, University of Coimbra, Portugal

in designing techniques for feature selection and extraction [9, 17]. As a result, there is growing interest in exploring alternative machine learning (ML) models, particularly deep learning (DL) models.

DL models, including convolutional neuronal networks (CNN), recurrent neural networks (RNN), and graph neural networks (GNN), have shown promise in capturing complex patterns within fMRI data [1, 3, 16]. These models enable brain state classification by leveraging whole-brain data [7, 19], or other data types derived from it like connectivity measures [12] or graph data [11, 18]. However, the requirement of large datasets required from these models poses a challenge in typical neuroimaging studies with limited participants. Transfer learning can address this challenge [8, 20] by leveraging large-scale fMRI projects such as the Human Connectome Project (HCP) [5] and BioBank [15], training DL models on extensive datasets and later generalizing them for use in common fMRI studies, thereby enhancing their applicability.

In this study, we present a novel approach to address the data limitations in fMRI studies. Our objective is to classify binary emotional arousal levels (High vs Low Arousal) using fMRI data obtained from a small sample of 16 participants. Drawing inspiration from EEGNet a compact CNN originally designed for electroencephalography (EEG) classification, we adapt the EEGNet architecture for fMRI data by introducing minor modifications to its structure. To accommodate the specific characteristics of fMRI data, we adjusted the original EEGNet structure and its derivative architectures [2]. Specifically, we reduced the size of the temporal filters to better handle the lower time dimensionality of fMRI data. Additionally, we transformed the 4D fMRI data into 2D by extracting the average time course of all voxels within predefined regions of interest (ROIs) based on an anatomical atlas. Furthermore, we explore the EEGNet architecture and develop our own networks that leverage the richer spatial information inherent in fMRI data.

2 Materials and methods

2.1 Preparation of fMRI data

The fMRI task involved three runs, each with 10 trials, following a block design. The blocks included fixation-cross, video watching, and self-evaluation. Videos were sourced from the CAAV dataset [4] and represented two levels of emotional arousal (high and low) in a 2:1 ratio. To create input samples for classification, the entire 15-second video block and an additional 5 seconds after were considered, covering the post-signal of the hemodynamic response. The mean time course of all voxels within predefined regions of interest (ROI) based on the brainnetone (BN) atlas [6] was extracted for each data block. The extracted time courses were normalized by subtracting the mean value of the previous 5 seconds. Data augmentation was performed using a sliding window of size 15, moving across the 20 seconds of data with a step size of one.

2.2 Classification approach

A leave-one-subject-out (LOSO) classification approach was employed, where each participant's data were used for testing, while the remaining participants' data were used for training. Participants were randomly split into training and validation sets, ensuring generalizability. This process was repeated ten times for each participant, and the models' performance was evaluated by calculating the mean balance accuracy across all participants. The state-of-the-art models, including EEGNet and its derivatives, are described elsewhere [2, 10]. Our own model, fMRINet, is summarized in Figure 1.

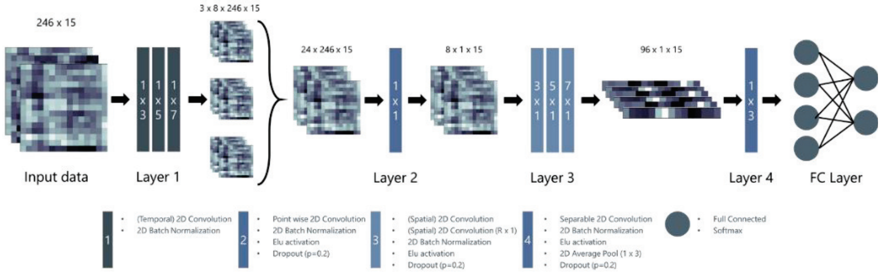

Fig. 1. Architecture summary of the fMRINet. This network comprises three layers convolution layers plus one fully connected layer. Details regarding each layer can be found in the bottom.

3 Preliminary Results

As shown in Table 1, our preliminary results demonstrate that all models achieved performances significantly above chance.

Table 1. Summary of the results obtained. The balanced accuracy is presented as the mean ± standard error. The p-value, obtained through a permutation test, indicates the statistical significance of the model's performance, specifically its performance above chance.

Model	Balance Accuracy	p-value
EEGNet	70.49±1.56	<0.01
EEGNetv2	72.53±1.56	<0.01
EEGNetv2.1	72.54±1.62	<0.01
EEGNeX	70.60±1.59	<0.01
fMRINet	**73.76±1.96**	**<0.01**
SVM	71.32±1.85	<0.01

These findings highlight the potential of leveraging EEGNet to build task-specific brain decoders, even with limited data availability. Importantly, our fMRINet model exhibited a slight performance advantage over the EEGNet models, suggesting the possibility of further enhancing these networks by incorporating the spatial characteristics inherent in fMRI data.

In future work, we aim to enhance these compact DL models and explore new approaches that utilize effective connectivity measures and graph neural networks (GNN). Our end objective is to develop low-complexity, easily interpretable models with biological significance.

References

1. Chan, H.P. et al.: Deep Learning in Medical Image Analysis. In: Advances in Experimental Medicine and Biology. (2020). https://doi.org/10.1007/978-3-030-33128-3_1.
2. Chen, X. et al.: Toward reliable signals decoding for electroencephalogram: A benchmark study to EEGNeX. (2022).
3. Cichy, R.M., Kaiser, D.: Deep Neural Networks as Scientific Models, (2019). https://doi.org/10.1016/j.tics.2019.01.009.
4. Di Crosta, A. et al.: The Chieti Affective Action Videos database, a resource for the study of emotions in psychology. Sci. Data. 7, 1, 32 (2020). https://doi.org/10.1038/s41597-020-0366-1.
5. Van Essen, D.C. et al.: The WU-Minn Human Connectome Project: An overview. Neuroimage. 80, (2013). https://doi.org/10.1016/j.neuroimage.2013.05.041.
6. Fan, L. et al.: The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. Cereb. Cortex. 26, 8, 3508–3526 (2016). https://doi.org/10.1093/cercor/bhw157.
7. Huang, X. et al.: Design of Deep Learning Model for Task-Evoked fMRI Data Classification. Comput. Intell. Neurosci. 2021, (2021). https://doi.org/10.1155/2021/6660866.
8. Kermany, D.S. et al.: Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell. 172, 5, (2018). https://doi.org/10.1016/j.cell.2018.02.010.
9. Kim, B., Oertzen, T. von: Classifiers as a model-free group comparison test. Behav. Res. Methods. 50, 1, (2018). https://doi.org/10.3758/s13428-017-0880-z.
10. Lawhern, V.J. et al.: EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. J. Neural Eng. 15, 5, (2018). https://doi.org/10.1088/1741-2552/aace8c.
11. Li, X. et al.: BrainGNN: Interpretable Brain Graph Neural Network for fMRI Analysis. Med. Image Anal. 74, (2021). https://doi.org/10.1016/j.media.2021.102233.
12. Meszlényi, R.J. et al.: Resting state fMRI functional connectivity-based classification using a convolutional neural network architecture. Front. Neuroinform. 11, (2017). https://doi.org/10.3389/fninf.2017.00061.
13. Showraki, M. et al.: Generalized Anxiety Disorder: Revisited. Psychiatr. Q. 91, 3, (2020). https://doi.org/10.1007/s11126-020-09747-0.
14. Smith, R. et al.: Neurocomputational mechanisms underlying emotional awareness: Insights afforded by deep active inference and their potential clinical relevance, (2019). https://doi.org/10.1016/j.neubiorev.2019.09.002.
15. Sudlow, C. et al.: UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. 12, 3, (2015). https://doi.org/10.1371/journal.pmed.1001779.
16. Tsuneki, M.: Deep learning models in medical image analysis, (2022). https://doi.org/10.1016/j.job.2022.03.003.

17. Vieira, S. et al.: Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications, (2017). https://doi.org/10.1016/j.neubiorev.2017.01.002.
18. Wang, L. et al.: Graph convolutional network for fmri analysis based on connectivity neighborhood. Netw. Neurosci. 5, 1, (2021). https://doi.org/10.1162/netn_a_00171.
19. Wang, X. et al.: Decoding and mapping task states of the human brain via deep learning. Hum. Brain Mapp. 41, 6, (2020). https://doi.org/10.1002/hbm.24891.
20. Wen, H. et al.: Transferring and generalizing deep-learning-based neural encoding models across subjects. Neuroimage. 176, (2018). https://doi.org/10.1016/j.neuroimage.2018.04.053.

Learning Postoperative Pain through Physiological Signals

Daniela Pais[0000−0003−2600−0757]*

Institute of Electronics and Informatics Engineering of Aveiro (IEETA), Department of Electronics, Telecommunications and Informatics (DETI), Intelligent Systems Associate Laboratory (LASI), University of Aveiro, 3810-193 Aveiro, Portugal
danielapais@ua.pt

Abstract. Choosing the appropriate treatment to manage postoperative pain depends on accurately assessing its intensity. However, the current assessment methods are subjective, discontinuous, and inadequate for evaluating the pain of patients unable to communicate verbally. Therefore, there is a need to develop an objective and continuous method that does not require patient reports. This work proposes to develop data science strategies based on physiological signals to monitor and manage postoperative pain more effectively. To this end, relevant physiological features that exhibit strong correlations with self-reported pain will be identified, enabling the prediction of postoperative pain intensity and detection of pain relief after medication, with the support of machine learning approaches.

Keywords: Data acquisition · Feature extraction · Machine learning · Pain assessment · Physiological signals · Postoperative pain

1 Introduction

After surgery, patients often experience postoperative pain [17]. Inadequate pain management has been associated with increased morbidity and mortality, adverse psychological effects, delayed recovery time, and higher medical costs [1, 8]. Therefore, managing postoperative pain effectively can reduce complications, shorten hospital stays, and, ultimately, promote better recovery of the patients.

An accurate assessment of pain intensity is essential for successful pain treatment [11]. Nowadays, pain is assessed using scales and questionnaires, relying,

* MAP Doctoral Program in Computer Science (MAPi), Universities of Minho, Aveiro and Porto
Advisors:
Raquel Sebastião, IEETA, DETI, LASI, University of Aveiro, Portugal (raquel.sebastiao@ua.pt)
Susana Brás, IEETA, DETI, LASI, University of Aveiro, Portugal (susana.bras@ua.pt)
Manuel Vico Ávalos, Tondela-Viseu Hospital Center & Faculty of Health Sciences, University of Beira Interior, Portugal (mvico@fcsaude.ubi.pt)

most of the time, on self-reporting. One widely used approach is the Numeric Rating Scale, where patients are asked to rate their pain level on a scale ranging from 0 to 10 (or 100), where 0 represents the absence of pain, and 10 (or 100) represents the worst pain imaginable [3]. While self-reporting is widely used as the most suitable method for pain assessment, when patients are unable to self-report due to cognitive conditions or communication limitations, healthcare professionals can rely on behavioral cues, such as facial expressions and body movements, and physiological measures, including heart rate, blood pressure, and respiratory rate, to identify distress in these individuals [14, 20].

Although these assessment tools used in routine clinical practice are validated, they present several limitations [22]. These include subjectivity in reports and difficulty in evaluating the subjective experience of pain from a third-person perspective, which can lead to potential observational bias [6]. Furthermore, assessment methods are not continuous and provide short-term measures, and behavior and physiological changes are not always specific to pain [9]. Therefore, the development of an objective continuous assessment method would be valuable for monitoring pain and guiding drug administration after surgery.

2 State of the art

Recent research has shown an association between pain perception and the autonomic nervous system (ANS). The ANS response to postoperative pain can be measured non-invasively, allowing for real-time monitoring of the balance in autonomic tone through physiological signals to guide clinical decision-making [7].

Several physiological indicators extracted from electrocardiogram (ECG), electrodermal activity (EDA), photoplethysmogram (PPG), pupillometry, and surface electromyography (sEMG) signals showed a correlation with postoperative pain [4, 5, 12, 16, 18]. Moreover, various physiological parameters have also been shown to be significantly different after pain treatment [10, 13, 19, 21]. Furthermore, a study demonstrated that employing a support vector machine (SVM) classifier with eight Heart Rate Variability (HRV) features resulted in accuracies of 67.03%, 84.79%, 76.18%, and 63.86%, respectively, when comparing baseline and increasing pain levels (PL1, PL2, PL3, PL4) [15]. In [2], EDA signals achieved accuracies of 86.0%, 70.0%, and 61.5% in classifying PL1, PL2, and PL4, respectively, by employing a random forest (RF) classifier. For the classification of PL3, a k-nearest-neighbor (kNN) classifier demonstrated an accuracy of 72.1%. Moreover, the systolic peak amplitude variation normalized by PPG AC amplitude showed an accuracy of 79.5% to distinguish between preoperative and postoperative conditions using logistic classification [23].

3 Methodology

In order to contribute with relevant advances in the physiological assessment of postoperative pain, this thesis aims to answer the following research questions:

- What is the relationship between human physiology and pain?
- Considering overlapped sliding windows, what is the optimal length to provide reliable physiological feature extraction?
- Which physiological features, or combinations, provide a better description of postoperative pain?
- Which features and Machine Learning (ML) algorithms will be best suited for pain level classification?

To address these questions, the following methodology will be pursued. The first phase corresponds to data collection from adult participants undergoing elective abdominal surgery at Centro Hospitalar Tondela-Viseu (CHTV), recruited on a volunteer base after written informed consent. The data includes physiological signals (such as ECG, EDA, EMG) obtained through minimally invasive techniques, self-reported pain assessments, as well as information regarding age, gender, surgery, and postoperative settings.

Following, signal preprocessing will be carried out, including signal filtering, and subsequent feature extraction. The effects of different anesthetic agent drugs on the collected physiological signals should be studied and taken into consideration when deriving pain indicators during the feature extraction process.

An initial study will investigate the effect of pain relief medication on the extracted features, comparing pain reports before and after pain management. Furthermore, the extracted features will be analyzed, both individually and in various combinations, which can involve features from the same physiological signal or features derived from different physiological signals, to determine their suitability as indicators of postoperative pain and identify which better correlates with self-reported pain. The goal is to derive multimodal Physiological Postoperative Pain indicators (mPPPi) and provide an ML model for pain intensity classification, exploring also the application of ML explainable models.

Considering the steps mentioned above, innovative solutions in applied data science according to data constraints expected in this type of data (class imbalance, noise, artifacts) will also be addressed. Furthermore, considering overlapping sliding windows, the optimal length to provide reliable feature extraction will be determined. Another goal involves providing graph-representation approaches for change detection (namely, changes that pain relief poses on mPPPi), allowing continuous monitoring and assessment of postoperative pain.

Finally, the ultimate goal is to develop a user interface to integrate the proposed methods for pain assessment and classification, which may also enable the evaluation of the achievements through a clinical perspective.

Acknowledgments This work was funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the PhD grant UI/BD/153374/2022 (D.P.), under the Scientific Employment Stimulus CEECIND/03986/2018 (R.S.) and CEECINST/00013/2021 (R.S.), within the R&D unit IEETA/UA (UIDB/00127/2020), and under the project EMPA (2022.05005.PTDC). This work is also funded by national funds, European Regional Development Fund, FSE through COMPETE2020, through FCT, in the

scope of the framework contract foreseen in the numbers 4, 5, and 6 of the article 23, of the Decree-Law 57/2016, of August 29, changed by Law 57/2017, of July 19 (S.B.).

References

1. Apfelbaum, J.L., Chen, C., Mehta, S.S., Gan, T.J.: Postoperative pain experience: results from a national survey suggest postoperative pain continues to be undermanaged. Anesthesia and Analgesia **97**(2), 534–540 (2003). https://doi.org/https://doi.org/10.1213/01.ANE.0000068822.10113.9E
2. Aqajari, S.A.H., Cao, R., Kasaeyan Naeini, E., Calderon, M.D., Zheng, K., Dutt, N., Liljeberg, P., Salanterä, S., Nelson, A.M., Rahmani, A.M.: Pain assessment tool with electrodermal activity for postoperative patients: Method validation study. JMIR Mhealth Uhealth **9**(5), e25258 (2021)
3. Breivik, H., Borchgrevink, P.C., Allen, S.M., Rosseland, L.A., Romundstad, L., Hals, E.K.B., Kvarstein, G., Stubhaug, A.: Assessment of pain. British Journal of Anaesthesia **101**(1), 17–24 (2008). https://doi.org/https://doi.org/10.1093/bja/aen103
4. Caton, L., Bolzon, M., Boschiero, D., Thayer, J.F., Gidron, Y.: Presurgical heart-rate variability strongly predicts less post-operative pain in patients with epilepsy. Journal of Psychosomatic Research **145**, 110421 (2021). https://doi.org/https://doi.org/10.1016/j.jpsychores.2021.110421
5. Charier, D., Vogler, M.C., Zantour, D., Pichot, V., Martins-Baltar, A., Courbon, M., Roche, F., Vassal, F., Molliex, S.: Assessing pain in the postoperative period: Analgesia nociception index™ versus pupillometry. British Journal of Anaesthesia **123**(4), e322–e327 (2019). https://doi.org/10.1016/j.bja.2018.09.031
6. Coghill, R.C.: Individual differences in the subjective experience of pain: New insights into mechanisms and models. Headache **50**(9), 1531—1535 (2010). https://doi.org/https://doi.org/10.1111/j.1526-4610.2010.01763.x
7. Cowen, R., Stasiowska, M.K., Laycock, H., Bantel, C.: Assessing pain objectively: the use of physiological markers. Anaesthesia **70**(7), 828—-847 (2015). https://doi.org/https://doi.org/10.1111/anae.13018
8. Gan, T.J.: Poorly controlled postoperative pain: prevalence, consequences, and prevention. Journal of Pain Research **10**, 2287–2298 (2017). https://doi.org/https://doi.org/10.2147/JPR.S144066
9. Hummel, P., van Dijk, M.: Pain assessment: current status and challenges. Seminars in Fetal and Neonatal Medicine **11**(4), 237–245 (2006). https://doi.org/https://doi.org/10.1016/j.siny.2006.02.004
10. Kantor, E., Montravers, P., Longrois, D., Guglielminotti, J.: Pain assessment in the postanaesthesia care unit using pupillometry: A cross-sectional study after standard anaesthetic care. European Journal of Anaesthesiology **31**(2), 91–97 (2014). https://doi.org/10.1097/01.EJA.0000434966.96165.c9
11. Kasaeyan Naeini, E., Shahhosseini, S., Subramanian, A., Yin, T., Rahmani, A.M., Dutt, N.: An edge-assisted and smart system for real-time pain monitoring. In: 2019 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). pp. 47–52 (2019). https://doi.org/10.1109/CHASE48038.2019.00023
12. Ledowski, T., Bromilow, J., Wu, J., Paech, M.J., Storm, H., Schug, S.A.: The assessment of postoperative pain by monitoring skin conductance: results of a prospective study. Anaesthesia **62**(10), 989–993 (2007)

13. Ling, P., Siyuan, Y., Wei, W., Quan, G., Bo, G.: Assessment of postoperative pain intensity by using photoplethysmography. Journal of Anesthesia **28**(6), 846–853 (2014). https://doi.org/https://doi.org/10.1007/s00540-014-1837-3
14. Maxwell, L.G., Fraga, M.V., Malavolta, C.P.: Assessment of pain in the newborn: An update. Clinics in Perinatology **46**(4), 693–707 (2019). https://doi.org/https://doi.org/10.1016/j.clp.2019.08.005
15. Naeini, E.K., Subramanian, A., Calderon, M.D., Zheng, K., Dutt, N., Liljeberg, P., Salantera, S., Nelson, A.M., Rahmani, A.M.: Pain recognition with electrocardiographic features in postoperative patients: Method validation study. Journal of Medical Internet Research **23**(5), e25079 (2021). https://doi.org/https://doi.org/10.2196/25079
16. Park, M., Kim, B.J., Kim, G.S.: Prediction of postoperative pain and analgesic requirements using surgical pleth index: a observational study. Journal of Clinical Monitoring and Computing **34**(3), 583–587 (2020). https://doi.org/https://doi.org/10.1007/s10877-019-00338-4
17. Pogatzki-Zahn, E.M., Segelcke, D., Schugb, S.A.: Postoperative pain—from mechanisms to treatment. Pain Reports **2**(2), e588 (2017). https://doi.org/https://doi.org/10.1097/PR9.0000000000000588
18. Schasfoort, F.C., Formanoy, M.A.G., Bussmann, J.B.J., Peters, J.W.B., Tibboel, D., Stam, H.J.: Objective and continuous measurement of peripheral motor indicators of pain in hospitalized infants: A feasibility study. Pain **137**(2), 323–331 (2008). https://doi.org/https://doi.org/10.1016/j.pain.2007.09.011
19. Sesay, M., Robin, G., Tauzin-Fin, P., Sacko, O., Gimbert, E., Vignes, J.R., Liguoro, D., Nouette-Gaulain, K.: Responses of heart rate variability to acute pain after minor spinal surgery: Optimal thresholds and correlation with the numeric rating scale. Journal of Neurosurgical Anesthesiology **27**(2), 148–154 (2015). https://doi.org/https://doi.org/10.1097/ana.0000000000000102
20. Severgnini, P., Pelosi, P., Contino, E., Serafinelli, E., Novario, R., Chiaranda, M.: Accuracy of critical care pain observation tool and behavioral pain scale to assess pain in critically ill conscious and unconscious patients: Prospective, observational study. Journal of Intensive Care **4**(68) (2016). https://doi.org/https://doi.org/10.1186/s40560-016-0192-x
21. Tapar, H., Suren, M., Karaman, S., Dogru, S., Karaman, T., Sahin, A., Altıparmak, F.: Evaluation of the perfusion index according to the visual analog scale in postoperative patients. Saudi Medical Journal **39**(10), 1006—-1010 (2018). https://doi.org/http://dx.doi.org/10.15537/smj.2018.10.23095
22. Williamson, A., Hoggart, B.: Pain: a review of three commonly used pain rating scales. Journal of Clinical Nursing **14**(7), 798–804 (2005). https://doi.org/https://doi.org/10.1111/j.1365-2702.2005.01121.x
23. Yang, Y.L., Seok, H.S., Noh, G.J., Choi, B.M., Shin, H.: Postoperative pain assessment indices based on photoplethysmography waveform analysis. Frontiers in Physiology **9** (2018). https://doi.org/https://doi.org/10.3389/fphys.2018.01199

Electrocardiogram for Biometric Recognition: Collectability, Stability and Application Challenges

Teresa M.C. Pereira[1,2,3]*

[1] Instituto de Engenharia Eletrónica e Informática de Aveiro (IEETA), Departamento de Electrónica, Telecomunicações e Informática (DETI), Laboratório Associado de Sistemas Inteligentes (LASI), Universidade de Aveiro, 3810-193 Aveiro, Portugal
[2] Instituto de Biofísica e Engenharia Biomédica (IBEB), Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal
[3] Aveiro Institute of Materials (CICECO), Departamento de Engenharia de Materiais e Cerâmica (DEMaC), Universidade de Aveiro, 3810-164, Aveiro, Portugal
teresamcp@ua.pt

Abstract. Innovative approaches to individuals recognition based on physiological and behavioral characteristics have emerged since surrogate representations of identity no longer suffice. This trend is encouraged by the increase of low-cost computational power, allowing these methodologies to be deployed with an efficiency that is expected to have a great impact on technology. Biometric recognition through Electrocardiogram (ECG) has recently seen progress, largely due to the uniqueness of the ECG signal. This thesis aims to develop a polymeric-based off-the-person ECG sensor to acquire ECG signals and propose new methods for biometric authentication and identification of individuals.

Keywords: Biometric Recognition · Electrocardiogram · ECG sensor · Data Mining · Machine Learning.

1 Introduction

A variety of societal facilities rely on recognition systems to protect and guard ourselves, our personal data, and our belongings. Several still depend on traditional methods, such as passwords [14]. However, their reuse across different services can lead to account takeovers. Hence, traditional representations of identity are no longer sufficient. This has led to a recent shift of interest towards the field of biometric recognition [14, 23]. The most common biometric trait is the fingerprint, nevertheless its acquisition can be easily circumvented by a skilled

* Doctoral Program in Informatics Engineering
Advisors:
Raquel Sebastião[1] (raquel.sebastiao@ua.pt)
Raquel C. Conceição[2] (rcconceicao@fc.ul.pt)
Vitor Sencadas[3] (vsencadas@ua.pt)

specialist. ECG has recently shown a great potential to be used as a biometric trait mainly due to its uniqueness and hidden nature [5, 6, 26]. Thus, this thesis aims to identify and authenticate individuals through their ECG signals, acquired with a polymeric-based ECG sensor, which will also be developed under the course of this PhD.

2 State-of-the-art

Biometrics is defined by the International Organization for Standardization (ISO) as “the automated recognition of individuals based on their behavioral and biological characteristics” [7], and is an increasingly growing multibillion-dollar market [12].

Distinctive features evaluated by biometrics, which are referred to as biometric traits, must have the following characteristics [21]: universality, uniqueness, stability, collectability, performance, acceptability, and circumvention. The advantages of using ECG as a biometric trait are its universality, hidden nature, and simple acquisition, while its disadvantages include the need for contact and the fact that ECG is variable over time [16].

A biometric recognition system (BRS) can be divided into data acquisition, feature extraction and classification. Concerning data acquisition, ECG can be acquired through “on-the-person” or “off-the-person” approaches. The former were mostly used in early biometrics; however, they present movement constraints due to the large number of electrodes required [3, 19, 21]. The latter have more recently emerged since they require few dry Ag/AgCl electrodes in contact with subject’s hands or fingers, being easily integrated in day-to-day objects [18, 23, 27]. Nevertheless, they can trigger allergic reactions [29], and so conductive polymers are being explored as ECG electrodes [10, 15, 22, 28]. Contactless approaches based on radar technologies are also emerging and their ability in ECG biometrics is being studied [20].

Despite ECG being often acquired from healthy subjects, at rest, some researchers have been exploring ECG biometrics for cardiac patients [1, 2] and during physical exercise [8, 9, 18], registering a decrease in system’s accuracy.

Regarding feature extraction, existing approaches can be fiducial, non-fiducial, or partially-fiducial. Although fiducial features result in higher accuracies [8, 13, 25], non-fiducial features [1, 3, 17] have reduced computational costs as they do not have to accurately detect fiducial points. Partially-fiducial approaches [4, 13] are significantly more uncommon, as they combine the use of non-fiducial approaches after the detection of fiducial points. Concerning the classification stage, either a classifier or a metric-based method (MBM) can be used. Classifiers, such as Support Vector Machines [11, 24], Nearest Neighbor [1, 4, 11] and Neural Networks [4, 27] are mostly used for identification tasks, whereas MBM using Euclidean [3, 17, 24, 25] and Cosine distances [3, 24] are used for both identification and authentication. Despite the different techniques explored in the literature, there is still no consensus on which ones lead to better performance.

3 Methodology

Considering the limitations and drawbacks from literature, the following research questions will be addressed:

1. Can a polymeric-based sensor be used for biometric recognition?
2. Which conditions for ECG acquisition should be considered when designing the data collection protocol for a BRS?
3. Which factors of data acquisition influence the intra- and inter-subject variability and how they impact the performance of a BRS?
4. Which machine learning approaches are most suitable for identification and authentication tasks? Simultaneously, a study on the most relevant features regarding the conditions of acquisition, and the limitation of classification methods will also be attained.

To answer these questions, the proposed methodology will be followed:

T1. State-of-the-art review, focusing on ECG sensors, ECG collection, Signal Processing, Data Mining, and Machine Learning. It should include a literature search, and a review of the selected studies, with their findings and limitations.

T2. Sensor development, in which different polymeric materials and textures will be explored. Then, the electrical conductivity of the polymer will be studied, as well as sensor geometry, and mechanical and electrical stability.

T3. The Data collection task will include the design of the protocols for three different sessions. The first protocol, called Daily Acquisition, consists of acquiring ECG data from the fingers several times using the self-developed sensor to assess the stability of ECG signals over time and analyze the influence of cardiac conditions, age, and gender. The second protocol consists of acquiring ECG signals during physical exercise to analyze the influence of exercise and posture. The third protocol involves acquiring ECG signals when inducing different emotional states through video elicitation.

T4. Devoted to Data mining, this task considers signal pre-processing, feature extraction, in which fiducial, non-fiducial, and partially-fiducial features will be explored, and feature transformation.

T5. Devoted to machine learning approaches. this task aims to train and refine several classification models for both the identification and authentication tasks. A comparative analysis to assess the performance of each classifier will focus on the strengths and weaknesses of each specific condition of acquisition and the features used.

T6. Final frameworks. This task aims to deliver final frameworks for identification and authentication, taking into account the optimal selection of data acquisition conditions, features, and classification methods. Reproducibility of results will be emphasized.

T7. Scientific paper and thesis writing. During the course of this Ph.D., it is expected to publish at least three journal papers, and disseminate the results at national and international events, ending with the writing and defense of the thesis.

Acknowledgments This work was funded by national funds through FCT — Fundação para a Ciência e a Tecnologia, I.P., under the PhD grant UI/BD/153605/2022 (T.M.C.P.), the Scientific Employment Stimulus CEECIND/03986/2018 (R.S.) and CEECINST/00013/2021 (R.S.), and within the RD units IEETA/UA UIDB/00127/2020, IBEB UIDB/00645/2020, and CICECO-Aveiro Institute of Materials UIDB/50011/2020, UIDP/50011/2020, and LA/P/0006/2020 (PIDDAC).

References

1. Becerra, M.A., Duque-Mejía, C., Hernandez, J.C.Z., Peluffo-Ordónez, D.H., Serna-Guarín, L., Delgado-Trejos, E., Revelo, J., Blanco, X.: Exploratory study of the effects of cardiac murmurs on electrocardiographic-signal-based biometric systems. Lecture Notes in Computer Science **11314**, 410–418 (2018). https://doi.org/10.1007/978-3-030-03493-1_43
2. Chen, M., Li, Y.F., Bao, S.D., Zhan, Y.J.: A comparative performance study of electrocardiogram-based human identity recognition. In: IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC. pp. 121–126 (2019). https://doi.org/10.1109/CSE/EUC.2019.00032
3. Coutinho, D.P., , Fred, L., Figueiredo, M.A.: ECG-based continuous authentication system using adaptive string matching. In: BIOSIGNALS 2011 - Proceedings of the International Conference on Bio-inspired Systems and Signal Processing. pp. 354–359 (2011). https://doi.org/10.5220/0003292003540359
4. Dar, M.N., Akram, M.U., Usman, A., Khan, S.A.: ECG biometric identification for general population using multiresolution analysis of DWT based features. In: 2015 Second International Conference on Information Security and Cyber Forensics (InfoSec). pp. 5–10 (2015). https://doi.org/10.1109/InfoSec.2015.7435498
5. Guven, G., Gurkan, H., Guz, U.: Biometric identification using fingertip electrocardiogram signals. Signal, Image and Video Processing **12**, 933–940 (2018). https://doi.org/https://doi.org/10.1007/s11760-018-1238-4
6. Huang, Y., Yang, G., Wang, K., Liu, H., Yin, Y.: Robust multi-feature collective non-negative matrix factorization for ECG biometrics. Pattern Recognition **123**, 108376 (2022). https://doi.org/https://doi.org/10.1016/j.patcog.2021.108376
7. ISO: Iso/iec 2382-37:2012 information technology — vocabulary — part 37: Biometrics, https://www.iso.org/standard/55194.html
8. Kim, K.S., Yoon, T.H., Lee, J.W., Kim, D.J., Koo, H.S.: A robust human identification by normalized time-domain features of electrocardiogram. Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference **2**, 1114–7 (2005). https://doi.org/10.1109/IEMBS.2005.1616615
9. Komeili, M., Louis, W., Armanfard, N., Hatzinakos, D.: On evaluating human recognition using electrocardiogram signals: From rest to exercise. In: 2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE) (2016). https://doi.org/10.1109/CCECE.2016.7726726
10. Kwon, S.H., Dong, L.: Flexible sensors and machine learning for heart monitoring. Nano Energy **102**, 107632 (2022). https://doi.org/https://doi.org/10.1016/j.nanoen.2022.107632
11. Lourenço, A., Silva, H., Fred, A.: ECG-based biometrics: A real time classification approach. In: 2012 IEEE International Workshop on Machine Learning for Signal Processing. pp. 1–6 (2012). https://doi.org/10.1109/MLSP.2012.6349735

12. MarketsAndMarkets: Biometric system industry worth $82.9 billion by 2027, https://www.marketsandmarkets.com/PressReleases/biometric-technologies.asp
13. Palaniappan, R., Krishnan, S.: Identifying individuals using ECG beats. In: 2004 International Conference on Signal Processing and Communications - SPCOM. pp. 569–572 (2004). https://doi.org/10.1109/SPCOM.2004.1458524
14. Pereira, T.M.C., Conceição, R.C., Sebastião, R.: Initial study using electrocardiogram for authentication and identification. Sensors **22**(6), 2202 (2022). https://doi.org/https://doi.org/10.3390/s22062202
15. Pereira, T.M.C., Conceição, R.C., Sencadas, V., Sebastião, R.: Biometric recognition: A systematic review on electrocardiogram data acquisition methods. Sensors **23**(3), 1507 (2023). https://doi.org/https://doi.org/10.3390/s23031507
16. Pinto, J.R., Cardoso, J.S., Lourenço, A.: Evolution, current challenges, and future possibilities in ECG biometrics. IEEE Access **6**, 34746–34776 (2018). https://doi.org/https://doi.org/10.1109/ACCESS.2018.2849870
17. Plataniotis, K.N., Hatzinakos, D., Lee, J.K.M.: ECG biometric recognition without fiducial detection. In: 2006 Biometrics Symposium: Special Session on Research at the Biometric Consortium Conference. pp. 1–6 (2006). https://doi.org/10.1109/BCC.2006.4341628
18. Ramos, M.S., Carvalho, J.M., Pinho, A.J., Brás, S.: On the impact of the data acquisition protocol on ECG biometric identification. Sensors **21**(14), 4645 (2021). https://doi.org/https://doi.org/10.3390/s21144645
19. Rathore, A.S., Li, Z., Jin, Z.: A survey on heart biometrics. ACM Computing Surveys **53**(6), 1–38 (2020). https://doi.org/10.1145/3410158
20. Rissacher, D., Galy, D.: Cardiac radar for biometric identification using nearest neighbour of continuous wavelet transform peaks. In: IEEE International Conference on Identity, Security and Behavior Analysis (ISBA 2015). pp. 1–6 (2015). https://doi.org/10.1109/ISBA.2015.7126356
21. Samarin, N., Sannella, D.: A key to your heart: Biometric authentication based on ECG signals. CoRR (2019), http://arxiv.org/abs/1906.09181
22. dos Santos Silva, A., Almeida, H., da Silva, H.P., Oliveira, A.: Design and evaluation of a novel approach to invisible electrocardiography (ECG) in sanitary facilities using polymeric electrodes. Scientific Reports **11**(62222) (2021). https://doi.org/https://doi.org/10.1038/s41598-021-85697-2
23. Silva, H., Lourenço, A., Fred, A.: In-vehicle driver recognition based on hand ECG signals. In: IUI'12: Proceedings of the 2012 ACM international conference on Intelligent User Interfaces. pp. 25–28 (2012). https://doi.org/https://doi.org/10.1145/2166966.2166971
24. da Silva, H.P., Fred, A., Lourenço, A., Jain, A.K.: Finger ECG signal for user authentication: usability and performance. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS). pp. 1–8 (2013). https://doi.org/10.1109/BTAS.2013.6712689
25. Singh, Y.N., Gupta, P.: ECG to individual identification. In: 2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems. pp. 1–8 (2008). https://doi.org/10.1109/BTAS.2008.4699343
26. Srivastva, R., Singh, A., Singh, Y.N.: Plexnet: A fast and robust ECG biometric system for human recognition. Information Sciences **558**, 208–228 (2021). https://doi.org/https://doi.org/10.1016/j.ins.2021.01.001
27. do Val Bento, N.F.A., Gamboa, H.: ECG biometrics using deep neural networks, http://hdl.handle.net/10362/75491

28. Warnecke, J.M., Ganapathy, N., Koch, E., Dietzel, A., Flormann, M., Henze, R., Deserno, T.M.: Printed and flexible ECG electrodes attached to the steering wheel for continuous health monitoring during driving. Sensors(Basel) **22**(11), 4198 (2022). https://doi.org/10.3390/s22114198
29. Wu, H., Yang, G., Zhu, K., Liu, S., Guo, W., Jiang, Z., Li, Z.: Materials, devices, and systems of on-skin electrodes for electrophysiological monitoring and human–machine interfaces. Advanced Science **8**(2), 2001938 (2020). https://doi.org/https://doi.org/10.1002/advs.202001938

AI in Natural Language Processing and Data

Clustering Massive, Noisy, and Unstructured Textual Streams

César Andrade*

Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal
up202101459@edu.fc.up.pt

Abstract. Many applications generate extensive short text as stream data. Traditional approaches often struggle to handle these sparse and high-dimensional data. Two main methods, similarity-based text stream clustering, and the probabilistic topic model approach, have been used in Short Text Stream Clustering (STSC) studies. However, these methods have limitations when dealing with massive, noisy, user-generated data, including misspellings, ambiguous abbreviations, and nonstandard shortening of words. This research aims to evaluate the performance of these clustering methods against other state-of-the-art approaches and to develop methods that integrate probabilistic and similarity-based approaches. The focus is on identifying the most efficient similarity-based methods, exploring various word representation techniques, and creating an integrated approach that effectively handles the challenges posed by short text data streams.

Keywords: Short Text Stream Clustering, Topic Model, Similarity

1 Introduction

The digital age has produced an unusual influx of text data from various sources such as social media and news sites. These data, often user-generated and unstructured, present notable challenges for automated text processing [1]. Classic methods need to be revised for these high-dimensional data, inspiring researchers to develop innovative techniques. These new methods aim to guide the complexity of these data, offering insights and opportunities in various fields [2].

Some studies in this field focus on Short Text Stream Clustering (STSC), which comprises two main divisions: similarity-based text stream clustering and the probabilistic topic model approach [3]. The first relies on the vector space model to represent documents and uses similarity metrics like cosine similarity to gauge the similarity between documents or clusters. The latter involves probabilistic topic models like Latent Dirichlet Allocation (LDA) [4] and its enhanced versions, representing a cluster as a cluster feature (CF) a vector and inferring

* Part of the MAP Doctoral Program in Computer Science. Supervised by Prof. João Gama from the Faculty of Economics and Prof. Rita P. Ribeiro from the Faculty of Sciences, University of Porto

a topic as a multinomial distribution over words. Despite recent advancements, short text stream clustering methods still struggle with high computational costs, especially when handling large-scale data [5], and exhibit a strong dependency on language or context/domain, which can limit their effectiveness [6].

This research aims to navigate these complexities by evaluating the effectiveness of combined similarity-based and probabilistic topic model methods in STSC, focusing mainly on short and noisy stream text data.

2 Related Work

The evolution of text clustering methods has been marked by significant advancements in both similarity-based methods and the development of word embeddings. Initially, similarity-based methods such as the Vector Space Model (VSM) were widely used, where documents were represented as vectors and similarity scores were calculated using metrics like cosine similarity [7–11] However, with the advent of neural networks, word embeddings became a powerful tool for capturing semantic and syntactic relationships between words, contributing significantly to the field [12–19].

The introduction of the Transformer architecture in 2017 marked a turning point in sequence modeling, offering a novel approach to handling text data [20–25]. This led to the development of various Transformer-based models such as BERT [26], and its variants [27–30], which have been widely applied in tasks involving short text stream clustering.

Despite these advancements in word representation and similarity methods, traditional clustering methods like K-Means and Hierarchical Clustering have remained relevant, being used independently of the word representation method [31–33]. Other techniques have also been explored, further enriching the landscape of text clustering methods [16, 34, 35]. This continuous evolution and interplay of methods have shaped the current state of text clustering, demonstrating the dynamic nature of this field.

Probabilistic models, particularly LDA-based models have been instrumental in short text stream clustering [3, 4, 36–46]. These models, while powerful, have certain limitations, especially when dealing with short and sparse text data. Several LDA variants and extensions have been proposed to address these issues. For instance, the Biterm Topic Model (BTM) [40] was designed to handle short texts by learning topics based on the aggregated biterms across the entire corpus, effectively mitigating the sparsity problem inherent in individual documents.

However, the dynamic nature of text streams presents additional challenges, as the topic distribution can change over time. To capture these temporal dynamics, models such as the Dynamic Topic Model (DTM) [37] and the Dynamic Mixture Model (DMM) [46] have been developed. These models extend the traditional LDA framework by allowing the topic distributions to evolve over time, making them more suitable for analyzing text streams. Despite these advancements, the field continues to evolve, with ongoing research aimed at further improving the performance and robustness of probabilistic models for STSC.

3 Research Questions

This research aims to push the boundaries of current text clustering methods, particularly those dealing with the complexity of short, noisy text streams commonly found in social media posts, product descriptions, news headlines, and question titles. The non-standardized nature and frequent truncations of these text streams have proven to be a significant challenge for traditional Natural Language Processing (NLP) approaches, with neither similarity-based nor probabilistic methods providing satisfactory solutions. We intend to contribute to further advancing the processing of short and noisy stream text data. We will focus on finding answers to the following research questions to attain that.

RQ1: Can a combination of Similarity-based and Probabilistic Topic Model methods effectively cluster such data? This potential synergy between the unsupervised clustering ability of Latent Dirichlet Allocation (LDA) [47] and the semantic learning capability of embeddings [48] could provide a robust solution to handle text data characterized by misspellings, ambiguous abbreviations, and non-standard shortenings of words.

RQ2: In the context of certain types of data, such as e-commerce, that naturally possess a hierarchical composition [49], would the joint use of hierarchical clustering methods prove beneficial for Short Text Stream Clustering (STSC)? The formation of micro-clusters by these methods could potentially reduce the scope of comparison in the final clustering step in massive data stream scenarios.

RQ3: What are the implications of recent advancements in word representation techniques on STSC methods? With the evolution from Bag of Words (BOW) to methods like Word2Vec, GloVe, and BERT transformers, how do these transitions influence STSC methods' effectiveness in dealing with text data with misspellings, ambiguous abbreviations, and non-standard shortenings of words, especially when working with Probabilistic and Similarity-based methods?

4 Proposed Methodology

Our goal is to explore how combining two methodological approaches, backed by state-of-the-art word representation techniques, can improve the performance in noise data processing.

We plan to utilize datasets sourced from various real-world applications such as the Brazilian electronic invoice project, Google News headlines, Twitter feeds, and StackOverflow question titles. Through these diverse datasets, we hope to examine the efficacy of various word representation models, including TFIDF, GloVe, Word2Vec, and BERT, with a specific emphasis on evaluating BERT's utility in a Similarity-based and LDA context. Additionally, we will assess a range of character and token-based similarity measures, such as Levenstein, Needleman-Wunsh, Jaro-Winkler, cosine, and n-gram. We are confident that our comprehensive and rigorous approach will yield insightful results, which we will evaluate using standard metrics, including Homogeneity, Completeness, Normalized Mutual Information, and Purity.

References

1. Gandomi, Amir and Haider, Murtaza: Beyond the hype: Big data concepts, methods, and analytics. In: *International Journal of Information Management*, volume 35, number 2, pp. 137-144 (2015)
2. Raghupathi, Wullianallur and Raghupathi, Viju: Big data analytics in healthcare: promise and potential. In: *Health Information Science and Systems*, volume 2, number 3 (2014)
3. Yin, J. and Zhang, W. and Chao, D. and Yu, X. and Liu, Z. and Wang, J.: Model-based clustering of short text streams. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2634-2642. ACM (2016)
4. Blei, D. M., Ng, A. Y., & Jordan, M. I.: Latent dirichlet allocation. In: *Journal of machine Learning research*, volume 3, number Jan, pp. 993–1022. (2003)
5. Fuentes-Pineda, Gibran and Meza-Ruiz, Ivan Vladimir: Topic Discovery in Massive Text Corpora Based on Min-Hashing. In: *arXiv preprint arXiv:1807.00938* (2018)
6. Santhanam, Sivasurya: Context based Text-generation using LSTM networks. In: *arXiv preprint arXiv:2005.00048* (2020)
7. Zhang, Tian and Ramakrishnan, Raghu and Livny, Miron: BIRCH: an efficient data clustering method for very large databases. In: *ACM sigmod record*, volume 25, number 2, pp. 103–114 (1996)
8. Carnein, Matthias and Trautmann, Heike: Optimizing data stream representation: An extensive survey on stream clustering algorithms. In: *Business & Information Systems Engineering*, volume 61, number 3, pp. 277–297 (2019)
9. Cao, Feng and Estert, Martin and Qian, Weining and Zhou, Aoying: Density-based clustering over an evolving data stream with noise. In: *Proceedings of the 2006 SIAM international conference on data mining*, pp. 328–339 (2006)
10. Lorbeer, Boris and Kosareva, Ana and Deva, Bersant and Softić, Dženan and Ruppel, Peter and Küpper, Axel. A-BIRCH: automatic threshold estimation for the BIRCH clustering algorithm. In: *INNS conference on Big Data*, pp. 169–178, 2016. Springer.
11. Shou, L., Wang, Z., Chen, K., & Chen, G.: Sumblr: continuous summarization of evolving tweet streams. In: *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 533–542. (2013)
12. Mikolov, T., Chen, K., Corrado, G., & Dean, J.: Efficient estimation of word representations in vector space. In: *arXiv preprint arXiv:1301.3781.* (2013)
13. Pennington, J., Socher, R., & Manning, C. D.: Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543. (2014)
14. Huang, J., Li, J., Chen, Y., Sun, J., & Shi, P.: Burst Hotspots Dynamic Detection and Tracking on Large-Scale Text Stream. In: *IEEE Access*, volume 7, pp. 30913–30924. (2019)
15. Liu, Wensong and Lin, Feng and Hu, Zhuqing and Zhang, Jinhui. Optimized Clustering based on Semantic Similarity of Components for Short Text. In: *2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*, pp. 1–6, 2019.
16. Yuan, Xu and Sun, Mingyang and Chen, Zhikui and Gao, Jing and Li, Peng: Semantic clustering-based deep hypergraph model for online reviews semantic classification in cyber-physical-social systems. In: *IEEE Access*, volume 6, pp. 17942–17951 (2018)

17. Paalman, Jasper and Mullick, Shantanu and Zervanou, Kalliopi and Zhang, Yingqian: Term based semantic clusters for very short text classification. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pp. 878–887 (2019)
18. Kenter, T., & De Rijke, M.: Short text similarity with word embeddings. In: *Proceedings of the 24th ACM international on conference on information and knowledge management*, pp. 1411–1420. (2015)
19. Ltaifa, Ibtihel Ben and Hlaoua, Lobna and Romdhane, Lotfi Ben. Hybrid Deep Neural Network-Based Text Representation Model to Improve Microblog Retrieval. *Cybernetics and Systems*, vol. 51, no. 2, pp. 115-139, 2020.
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I.: Attention is all you need. In: *Advances in neural information processing systems*, volume 30. (2017)
21. Bodrunova, S. S., Orekhov, A. V., Blekanov, I. S., Lyudkevich, N. S., & Tarasov, N. A.: Topic detection based on sentence embeddings and agglomerative clustering with markov moment. In: *Future Internet*, volume 12, number 9, pp. 144. MDPI (2020)
22. Silva, Diego F. and Silva, Alcides M. e and Lopes, Bianca M. and Johansson, Karina M. and Assi, Fernanda M. and de Jesus, Júlia T. C. and Mazo, Reynold N. and Lucrédio, Daniel and Caseli, Helena M. and Real, Livy. Named Entity Recognition for Brazilian Portuguese Product Titles. In: *Journal Article*, pp. 526-541, 2021.
23. Asyaky, Muhammad Sidik and Mandala, Rila. Improving the Performance of HDBSCAN on Short Text Clustering by Using Word Embedding and UMAP. In: *Proceedings - 2021 8th International Conference on Advanced Informatics: Concepts, Theory, and Application, ICAICTA 2021*, 2021.
24. Harsh Sakhrani, Saloni Parekh, and Pratik Ratadiya. "Transformer-based Hierarchical Encoder for Document Classification." In *2021 International Conference on Data Mining Workshops (ICDMW)*, pp. 852–858, 2021.
25. Wang, Quan and Mao, Zhendong and Wang, Bin and Guo, Li. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2724-2743, 2017.
26. Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv preprint arXiv:1810.04805*, 2018.
27. Sanh, Victor and Debut, Lysandre and Chaumond, Julien and Wolf, Thomas. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 2019.
28. Souza, Fábio and Nogueira, Rodrigo and Lotufo, Roberto. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: *Lecture Notes in Computer Science*, vol. 12319 LNAI, pp. 403-417, 2020.
29. Liu, Yinhan and Ott, Myle and Goyal, Naman and Du, Jingfei and Joshi, Mandar and Chen, Danqi and Levy, Omer and Lewis, Mike and Zettlemoyer, Luke and Stoyanov, Veselin. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019.
30. Denisov, Sergej and Bäumer, Frederik S. How to Improve E-commerce Search Engines? Evaluating Transformer-Based Named Entity Recognition on German Product Datasets. In: *Communications in Computer and Information Science*, vol. 1486 CCIS, pp. 353-366, 2021.
31. Yin, Hui and Song, Xiangyu and Yang, Shuiqiao and Huang, Guangyan and Li, Jianxin. Representation Learning for Short Text Clustering. In: *International Conference on Web Information Systems Engineering*, pp. 321–335, 2021. Springer.

32. Molina, Roberto and Hasperué, Waldo and Villa Monte, Augusto. D3CAS: Distributed Clustering Algorithm Applied to Short-Text Stream Processing. In: *Argentine Congress of Computer Science*, pp. 211–220, 2018. Springer.
33. Geng, Fei and Liu, Qilie and Zhang, Ping: A time-aware query-focused summarization of an evolving microblogging stream via sentence extraction. In: *Digital Communications and Networks*, volume 6, number 3, pp. 389–397 (2020)
34. Haodi Zhong, Grigorios Loukides, and Solon P Pissis. "Clustering sequence graphs." *Data & Knowledge Engineering*, vol. 138, p. 101981, 2022.
35. Peng, Zijun and Xin, Guodong and Wei, Yuliang and Wang, Wei and Wang, Bailing and Wang, Lianhai. Short Text Clustering Enhanced by Semantic Matching Model. In: *2019 2nd International Conference on Information Systems and Computer Aided Education (ICISCAE)*, pp. 480–484, 2019. IEEE.
36. Yin, J., & Wang, J.: A model-based approach for text clustering with outlier detection. In: *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pp. 625–636. IEEE (2016)
37. Blei, D. M., & Lafferty, J. D.: Dynamic topic models. In: *ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning*, volume 2006, pp. 113 – 120. (2006)
38. Kumar, J., Shao, J., Uddin, S., & Ali, W.: An online semantic-enhanced Dirichlet model for short text stream clustering. In: *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 766–776. (2020)
39. Chen, J., Gong, Z., & Liu, W.: A nonparametric model for online topic discovery with word embeddings. In: *Information Sciences*, volume 504, pp. 32-47. 2019.
40. Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. "A biterm topic model for short texts." In *Proceedings of the 22nd international conference on World Wide Web*, pp. 1445–1456, 2013.
41. Chen, J., Gong, Z., & Liu, W.: A Dirichlet process biterm-based mixture model for short text stream clustering. In: *Applied Intelligence*, volume 50, number 5, pp. 1609-1619. 2020.
42. Shuiqiao Yang, Guangyan Huang, Bahadorreza Ofoghi, and John Yearwood. "Short text similarity measurement using context-aware weighted biterms." *Concurrency and Computation: Practice and Experience*, vol. 34, no. 8, p. e5765, 2022.
43. Lu-yao Xie, Lu-Xia Wang, Heng-Yang Lu, Ning Li, and Chong-Jun Wang. "Topics may Evolve: Using Complaint Data for Analysis." In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 1296–1303, 2017.
44. Rakib, M. R. H., & Asaduzzaman, M.: Fast Clustering of Short Text Streams Using Efficient Cluster Indexing and Dynamic Similarity Thresholds. In: *CoRR*, volume abs/2101.08595. (2021)
45. Jiajia Huang, Min Peng, Pengwei Li, Zhiwei Hu, and Chao Xu. "Improving biterm topic model with word embeddings." *World Wide Web*, vol. 23, no. 6, pp. 3099–3124, 2020.
46. Jianhua Yin and Jianyong Wang. "A dirichlet multinomial mixture model-based approach for short text clustering." In *Proceedings of the 20th ACM SIGKDD Inter. Conf. on Knowledge discovery and data mining*, pp. 233–242, 2014.
47. Michael Heck, S. Sakti and Satoshi Nakamura. Unsupervised Linear Discriminant Analysis for Supporting DPGMM Clustering in the Zero Resource Scenario. (2016).
48. Jianqiang Ma, Çagri Çöltekin and E. Hinrichs. Learning Phone Embeddings for Word Segmentation of Child-Directed Speech.(2016).
49. Houmin Yan, C. Sriskandarajah, S. Sethi and Xiaohang Yue. Supply-chain redesign to reduce safety stock levels: sequencing and merging operations. IEEE Trans. Engineering Management, (2002).

Data Quality, Data Balance and Data Documentation: a framework

Marco Rondina*[0009−0008−8819−3623]

Politecnico di Torino, marco.rondina@polito.it

Abstract. To promote the responsible development and use of artificial intelligence (AI), principles of trustworthiness, accountability and fairness should be followed. These principles can be threatened by various ethical challenges. Several approaches can be used, such as working at the data level, as data is one of the key elements of the AI pipeline. However, ensuring high data quality alone is not sufficient to avoid all ethical concerns. Expanding data quality frameworks to include data balance and data documentation can be helpful in addressing critical aspects of ethical considerations related to AI systems. The proposed framework introduces additional quality measures, such as the assessment of data balance and the quality of data documentation. The former can be useful to identify risks of disproportionate treatment of different groups based on their protected characteristics. The latter emphasises the importance of documenting datasets, making them more transparent and accountable. By integrating these measures into the development pipeline through appropriate data labels, we aim to empower practitioners to build more responsible systems. We outline future research directions for automating these metric evaluation processes.

Keywords: Data quality · data balance · data documentation · algorithmic bias · AI trustworthiness · AI accountability · AI fairness.

1 Introduction

Artificial intelligence (AI) has made significant advances in recent years, and concerns have arisen about its irresponsible development and use. Principles of trustworthiness, accountability and fairness should be followed to develop responsible AI systems. However, these principles can be threatened by various ethical challenges. These ethical problems may have roots in different portions of the AI pipeline, including data. A major concern is the use of poor quality, biased and poorly documented training data. Data quality plays a critical role in optimizing software performance, but ensuring data quality alone does not address all the ethical concerns associated with the use of AI systems. The focus of this research is to create a theoretical framework useful to cover the most relevant ethical challenges of an AI system [9]. By exploring these aspects from

* PhD program in Computer and Control Engineering. Advisors: Juan Carlos De Martin, Politecnico di Torino; Antonio Vetrò, Politecnico di Torino.

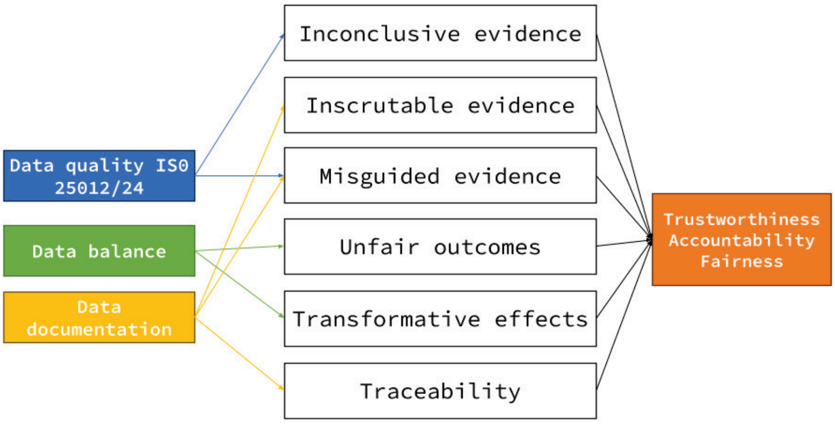

Fig. 1: The data dimensions (on the left) have an impact on the various ethical challenges (in the middle), with consequences for the desiderable proprieties of AI systems (on the right).

the data perspective, we aim to improve the trustworthiness, the accountability and the fairness of AI systems. This framework can also find practical application in the realisation of informative data labels.

2 State of the art

Responsibility in the selection, creation and adoption of datasets is at the heart of mitigating some problems, as faulty data leads to faulty AI models, which then produce undesirable outcomes. Therefore, researchers have started to explore ways to assess the quality, balance and provenance of datasets. Various works highlighted the importance of this topic, as algorithmic results are often data-driven [3,2]. In addition, noisy/inaccurate data reduce the validity of results [7,5], and data quality issues cause downstream effects (*cascades*) that are both common and avoidable [11]. A growing body of literature has explored how to make the intrinsic properties of datasets emerge. Gebru et al. proposed a list of questions useful to guide the writing of documentation by dataset creators, borrowing the concept of ”datasheet” from electronics. Holland et al. [4] proposed a diagnostic framework, as if it were a food label of a dataset. Bender et al. [1] focused on *data statements*, which can mitigate problems related to exclusion and bias in language technology. All the proposals presented above focused on data, but the question of what information is useful to bring out the particularities of a system is also relevant for models [8,10] or rankings [13,14].

Table 1: Data quality measures (ISO/IEC 25024).

Quality ID	**Characteristic**	**Description**
Acc-I-4	Accuracy	Ratio of outliers
Com-I-1 (adapted)	Completeness	Completeness of data items of a record within a data file
Com-I-5	Completeness	False completeness of records within a data file
Con-I-2 (adapted)	Consistency	Consistency of data format of the same data item
Con-I-3	Consistency	Risk of having inconsistency due to duplication
Con-I-4 (adapted)	Consistency	Degree to which the elements of the architecture have a correspondence in referenced architecture elements

Table 2: Imbalance indexes. m: number of classes; f_i: relative frequency of i.

Index	**Formula (normalized)**	**Notes**
Gini	$G_n = \frac{m}{m-1} \cdot \left(1 - \sum_{i=m}^{m} f_i^2\right)$	Measure of heterogeneity
Shannon	$S = -\left(\frac{1}{lnm}\right) \sum_{i=1}^{m} f_i ln f_i$	Measure of species diversity
Simpson	$D = \frac{1}{m-1} \cdot \left(\frac{1}{\sum_{i=1}^{m} f_i^2} - 1\right)$	Probability measure that two individuals randomly selected from a sample belong to the same class
Inverse Imbalance Ratio	$IR = \frac{\{min(f_i,\ldots,m)\}}{\{max(f_i,\ldots,m)\}}$	Ratio between the lowest and the highest frequency

3 Methodology

We aimed to investigate some relevant dimensions of data measurements, specifically addressing data quality, sensitive attribute imbalance and completeness of documentation. These factors can significantly impact the effectiveness and ethical implications of AI systems. We mapped these dimensions of data measurements with the ethical challenges identified by Mittelstadt et al. [9], as shown in Figure 1.

Firstly, we propose to evaluate data quality using ISO 25024 standard [6], as shown in Table 1. This includes assessments of accuracy, completeness, and consistency of the provided data. Second, we propose to measure data balance by using some heterogeneity metrics that have already been validated in the literature and could potentially be extended, such as the Gini index, the Shannon diversity index, the Simpson diversity index, and the inverse imbalance ratio [12]. The metrics are shown in the Table 2. An unbalanced dataset may perpetuate existing patterns of discrimination, potentially leading to unfair treatment of marginalized groups. Finally, we propose to assess the completeness of the documentation provided with each dataset. We integrate the metrics described above with the completeness metric provided by the Documentation Test Sheet[1].

[1] The Documentation Test Sheet is described in the paper ‘Completeness of Datasets Documentation on ML/AI repositories: an Empirical Investigation’, EPIA23, ERAI.

References

1. Bender, E.M., Friedman, B.: Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Transactions of the Association for Computational Linguistics **6**, 587–604 (2018). https://doi.org/10.1162/tacl_a_00041
2. Chen, Z., Zhang, J.M., Sarro, F., Harman, M.: A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers (Feb 2023). https://doi.org/10.48550/arXiv.2207.03277
3. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 329–338. FAT* '19, Association for Computing Machinery, New York, NY, USA (Jan 2019). https://doi.org/10.1145/3287560.3287589
4. Holland, S., Hosny, A., Newman, S., Joseph, J., Chmielinski, K.: The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. arXiv:1805.03677 [cs] (2018). https://doi.org/10.48550/arXiv.1805.03677
5. Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., Narayanan, A.: The Worst of Both Worlds: A Comparative Analysis of Errors in Learning from Data in Psychology and Machine Learning. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. pp. 335–348. AIES '22, Association for Computing Machinery, New York, NY, USA (Jul 2022). https://doi.org/10.1145/3514094.3534196
6. International Organization for Standardization: Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Measurement of data quality, https://www.iso.org/standard/35749.html
7. Kilkenny, M.F., Robinson, K.M.: Data quality: "Garbage in – garbage out" **47**(3), 103–144 (Sep 2018). https://doi.org/10.1177/1833358318774357
8. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model Cards for Model Reporting. In: Proc. of the 2019 Conf. on Fairness, Accountability and Transparency. pp. 220–229. FAT* '19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3287560.3287596
9. Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: The ethics of algorithms: Mapping the debate. Big Data & Society **3**(2) (Dec 2016). https://doi.org/10.1177/2053951716679679
10. Richards, J., Piorkowski, D., Hind, M., Houde, S., Mojsilović, A.: A Methodology for Creating AI FactSheets. arXiv:2006.13796 [cs] (2020). https://doi.org/10.48550/arXiv.2006.13796
11. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., Aroyo, L.M.: "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In: Proc. of the 2021 CHI Conf. on Hum. Factors in Comput. Syst. pp. 1–15. CHI '21, ACM, New York, NY, USA (2021). https://doi.org/10.1145/3411764.3445518
12. Vetrò, A., Torchiano, M., Mecati, M.: A data quality approach to the identification of discrimination risk in automated decision making systems. Government Information Quarterly **38**(4), 101619 (Oct 2021). https://doi.org/10.1016/j.giq.2021.101619
13. Yang, K., Stoyanovich, J., Asudeh, A., Howe, B., Jagadish, H.V., Miklau, G.: A Nutritional Label for Rankings. In: Proc. 2018 Int. Conf. on Manage. of Data. pp. 1773–1776 (2018). https://doi.org/10.1145/3183713.3193568

14. Zehlike, M., Yang, K., Stoyanovich, J.: Fairness in Ranking: A Survey (2021). https://doi.org/10.48550/arXiv.2103.14000

Estimating the Density Ratio with a ReLU Induced Tessellation*

Michał Lewandowski**[1,2]

[1] Johannes Kepler University (JKU), Altenberger Straße 66b, 4040 Linz, Austria
[2] Software Competence Center Hagenberg (SCCH), Softwarepark 32a, 4232 Hagenberg, Austria
michal.lewandowski@scch.at

Abstract. Density ratio estimation, crucial in machine learning, is hindered by challenges such as high-dimensionality data. Current methods struggle with complex data or have high computational costs. In this work, we try to tackle some of the existing issues using disjoint linear regions that are created during the learning process of neural networks with ReLU activation function, providing an innovative approach to estimate density ratios, with applications in unsupervised domain adaptation problems.

Keywords: density ratio estimation · ReLU neural networks

1 Introduction

The ratio of probability density functions is an important ingredient of many statistical and machine learning problems, such as anomaly detection [4], conditional density estimation [21], or unsupervised domain adaptation [5]. There exist numerous ways of estimating the density ratio in low dimensional problems [19, 20], but its estimation in high dimensions proves difficult [7]. To address this issue, we propose a method based on neural networks with ReLU activation function. We leverage the fact that a ReLU neural network during the training process tessellates the input space into many disjoint cells, called linear regions [13, 16]. We propose to construct the density ratio estimator by counting observations x, x' from the source and target domains in each linear region, and then take a (smoothed) ratio of counts from the target domain over the counts of observations from the source domain. Finally, we evaluate the efficacy of our approach on downstream tasks that require access to accurate density ratios, such as unsupervised domain adaptation (UDA). Contributions:

- We introduce an algorithm for estimating density ratios using the underlying geometry of the input space created by a ReLU neural network.
- We compare our approach with relevant baselines on tasks that require access to accurate density ratios, and show its advantages in high dimensions.

* PhD Student in Artificial Intelligence at JKU, supported by SCCH

** Supervised by Priv-Doz. Dr. Bernhard A.Moser, Senior Researcher at Inst. of Signal Processing at JKU

Related work. Explicit density estimates of $\hat{p}_{\mathcal{S}}(x)$ and $\hat{p}_{\mathcal{T}}(x)$ often perform poorly [19], inspiring direct ratio estimation techniques like Kernel Mean Matching (KMM) [5] and KLIEP [19]. However, these methods assume overlapping distributions, potentially unrealistic in real-world scenarios. Solutions have been proposed, including *telescoping density-ratio estimation* (TRE) for highly dissimilar densities [15]. Deep learning, using Convolutional Neural Networks (CNN) with uLSIF criterion [8], has also been applied [14]. Space tessellation methods also exist, with Veronoi [12] and Delaunay [1] tessellations used to estimate probability density functions. Our work uses a combination of tessellation and deep learning, akin to [12] and [14].

Preliminaries. A ReLU neural network, $\mathcal{N} : \mathcal{X} \to \mathcal{Y}$, consists of alternating ReLU functions (defined as $\mathrm{ReLU}(x) = \max(x, 0)$, and denoted with $\sigma(x)$) and affine functions with weights W_k and biases b_k at layer k. Let us denote the total number of hidden neurons in $\mathcal{N}$ by N. An input $x \in \mathcal{X}$, propagated through $\mathcal{N}$, generates non-negative activation values. An *activation pattern* is the binarization of these values, $\pi_{\mathcal{N}} : \mathcal{X} \to \{0,1\}^N$, and represents an element in a binary hypercube $\mathcal{H}_N := \{0,1\}^N$. Hamming distance, $d_H(u,v) := |\{u_i \neq v_i \text{ for } i = 0, \ldots, N\}|$, measures the difference between $u, v \in \mathcal{H}_N$. [17] proved that there exists a bijective mapping between activation patterns and equivalence classes, $[x]_{\mathcal{N}} := \{z \in \mathcal{X} |\, z \sim_{\mathcal{N}} x\}$, under the relation $x \sim_{\mathcal{N}} y \iff d_H(\pi_{\mathcal{N}}(x), \pi_{\mathcal{N}}(y)) = 0$. We call equivalence classes *linear regions*, and their collection a *tessellation*. The activation space, $(\mathcal{H}_N, d)$, is a metric space.

We state the problem as follows. Given two sets of observed samples x with labels y from source and target domains, $\mathcal{D}_{\mathcal{S}} = \{x_i, y_i\}_{i=1}^{n_{\mathcal{S}}} \sim p_{\mathcal{S}}(x,y)$ and $\mathcal{D}_{\mathcal{T}} = \{x_j, y_j\}_{j=1}^{n_{\mathcal{T}}} \sim p_{\mathcal{T}}(x,y)$, we wish to estimate the ratio of their underlying probability densities $w(x) = p_{\mathcal{S}}(x)/p_{\mathcal{T}}(x)$.

2 ReLU-Tessellation based Estimator of Density Ratio

We introduce counting measures for points from source and target domains in linear regions E_i, respectively $\mu(E_i) := |\{x \in E_i : x \sim p_{\mathcal{S}}\}|$, $\nu(E_i) := |\{x' \in E_i : x' \sim p_{\mathcal{T}}\}|$, and denote the density ratio estimator by $\widehat{w}_i := \nu(E_i)/\mu(E_i)$. To avoid issues with the division by zero, we use the smoothed counting measure $\mu_s(E_i) := \mu(E_i) + 1$, and refer to $\widehat{w}_i$ as smoothed weight. To improve accuracy of $\widehat{w}_i$, we propose to control the volume of linear regions by adding a regularization during the learning process $\frac{1}{n_{\mathcal{S}}} \sum_{i=1}^{n_{\mathcal{S}}} \mathrm{CE}\left(f\left(x_i\right), y_i\right) + \lambda\, \mathrm{MMR}\left(x_i\right)$, where $\lambda \in \mathbb{R}_+$ is a regularization parameter, CE is the usual cross-entropy loss, and MMR is a Maximum Margin Regularizer (Section 4, [3]).

Domain Adaptation. In UDA problems, source and target risk can be related as $\mathbb{E}_{(x,y) \sim \mathcal{D}_{\mathcal{T}}}\left[\ell\left(h\left(x\right), y\right)\right] = \mathbb{E}_{(x,y) \sim \mathcal{D}_{\mathcal{S}}}\left[\ell\left(h\left(x\right), y\right) w(x,y)\right]$. Joint distributions $p(x,y)$ are rarely tractable, hence one usually decomposes them as $p(x,y) = p(x|y)p(y)$, or $p(x,y) = p(y|x)p(x)$. We assume the latter with $p(y'|x') = p(y|x)$ and $p(x') \neq p(x)$, so called *covariate shift* [9]. To estimate $\widehat{w}_i$, we assign points x and x' to linear region E_i using their activation patterns $\pi_{\mathcal{N}}(x), \pi_{\mathcal{N}}(x')$, and estimate the target risk as $\mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathcal{T}}}\left[\ell\left(h(x),y\right)\right] \approx \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathcal{S}}\cap E_i}\left[\ell\left(h(x),y\right)\widehat{w}_i\right].$

3 Experimental protocol

We evaluate the proposed method on both synthetic and real data. We start with the synthetic data: multivariate gaussian distributions, where we can analytically calculate true density ratio $w(x)$ at any given point $x \in supp(\mathcal{D}_{\mathcal{S}}) \cap supp(\mathcal{D}_{\mathcal{T}})$. Then, we move on to two pairs of image datasets: MNIST [11] with USPS [6] and CIFAR-10 [10] with STL-10 [2], where the true density ratio is unknown.

Toy example. We start experiments with a fixed number of observations $(n_{\mathcal{S}}, n_{\mathcal{T}})$ sampled from two different multivariate gaussian distributions. At this stage, we are interested in assessing the accuracy of our method compared to other known density ratio estimation methods. The learning problem is constructed from both $\mathcal{D}_{\mathcal{S}}$ and $\mathcal{D}_{\mathcal{T}}$, and is a two-class classification problem. We investigate two scenarios: one with (1) unbalanced classes, where $n_{\mathcal{S}} = 50, n_{\mathcal{T}} = 100$, and $n_{\mathcal{S}} = 100, n_{\mathcal{T}} = 50$, and one with (2) balanced out classes with $n_{\mathcal{S}} = n_{\mathcal{T}} = 50$ for all datasets. We consider:

Table 1. Gaussian datasets. We provide moments μ and Σ for $N(\mu, \Sigma)$.

		"Shifting"	"Shrinking"	"Magnifying"	"Rotating"
μ	$x_{\mathcal{S}}$	$(0 \quad 0)^T$	$(0 \quad 0)^T$	$(0 \quad 0)^T$	$(1 \quad 0)^T$
	$x_{\mathcal{T}}$	$(0 \quad 3)^T$	$(0 \quad 3)^T$	$(0 \quad 3)^T$	$\frac{1}{\sqrt{2}}(1 \quad 1)^T$
Σ	$x_{\mathcal{S}}$	$\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$	$4\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$	$\frac{1}{4}\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$	$\begin{pmatrix}4 & 0\\0 & 1\end{pmatrix}$
	$x_{\mathcal{T}}$	$\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$	$\frac{1}{4}\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$	$4\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$	$\frac{1}{2}\begin{pmatrix}5 & 3\\3 & 5\end{pmatrix}$

This is an instructive example as we know explicitly the density ratio value $r(x)$. We propose to measure the performance of our method with mean squared error (MSE) from the true density ratio, and then compare it with the MSEs of competing methods [18]. To remove stochastic effects caused by network initialization we repeat our experiments over a defined set of random seeds. As the result the average over this repetitions is provided.

High-dimensional example. We use previously described image datasets, learn a ReLU neural network on a source domain, and investigate its accuracy on the target domain through domain adaptation as described previously. Then, we measure classification error on the target domain. This procedure is, again, averaged over a defined set of random seeds.

References

1. Belov, V., Marik, R.: Tessellation-based kernel density estimation. In: Proceedings of the 2021 4th International Conference on Algorithms, Computing and Artificial Intelligence. ACAI '21, Association for Computing Machinery, New York, NY, USA (2022)
2. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 15, pp. 215–223. Fort Lauderdale, FL, USA (11–13 Apr 2011)
3. Croce, F., Andriushchenko, M., Hein, M.: Provable robustness of relu networks via maximization of linear regions. In: Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 89, pp. 2057–2066. PMLR (16–18 Apr 2019)
4. Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., Kanamori, T.: Statistical outlier detection using direct density ratio estimation. Knowledge and Information Systems **26**(2), 309–336 (2011)
5. Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., Smola, A.: Correcting sample selection bias by unlabeled data. In: Advances in Neural Information Processing Systems. vol. 19. MIT Press (2007)
6. Hull, J.J.: A database for handwritten text recognition research. IEEE Transactions on Pattern Analysis and Machine Intelligence **16**(5), 550–554 (1994)
7. Izbicki, R., Lee, A., Schafer, C.: High-dimensional density ratio estimation with extensions to approximate likelihood computation. In: Artificial Intelligence and Statistics. pp. 420–429. PMLR (2014)
8. Kanamori, T., Hido, S., Sugiyama, M.: Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems. vol. 21. Curran Associates, Inc. (2009)
9. Kouw, W.M.: An introduction to domain adaptation and transfer learning. CoRR **abs/1812.11806** (2018)
10. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep. (2009)
11. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
12. Loog, M.: Nearest neighbor-based importance weighting. In: 2012 IEEE International Workshop on Machine Learning for Signal Processing. pp. 1–6 (2012)
13. Montúfar, G.F., Pascanu, R., Cho, K., Bengio, Y.: On the number of linear regions of deep neural networks. In: Advances in Neural Information Processing Systems. vol. 27 (2014)
14. Nam, H., Sugiyama, M.: Direct density ratio estimation with convolutional neural networks with application in outlier detection. IEICE Transactions on Information and Systems **E98.D**(5), 1073–1079 (2015)
15. Rhodes, B., Xu, K., Gutmann, M.U.: Telescoping density-ratio estimation. In: Advances in Neural Information Processing Systems. vol. 33, pp. 4905–4916 (2020)
16. Serra, T., Tjandraatmadja, C., Ramalingam, S.: Bounding and counting linear regions of deep neural networks. In: International Conference on Machine Learning. pp. 4558–4566 (2018)
17. Shepeleva, N., Zellinger, W., Lewandowski, M., Moser, B.: Relu code space: A basis for rating network quality besides accuracy. ICLR, NAS workshop (2020)

18. Sugiyama, M., Kawanabe, M., Chui, P.L.: Dimensionality reduction for density ratio estimation in high-dimensional spaces. Neural Networks **23**(1), 44–59 (2010)
19. Sugiyama, M., Nakajima, S., Kashima, H., Bünau, P.v., Kawanabe, M.: Direct importance estimation with model selection and its application to covariate shift adaptation. In: Proceedings of the 20th International Conference on Neural Information Processing Systems. p. 1433–1440. NIPS'07, Red Hook, NY, USA (2007)
20. Sugiyama, M., Suzuki, T., Kanamori, T.: Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. Annals of the Institute of Statistical Mathematics **64**(5), 1009–1044 (2012)
21. Sugiyama, M., Takeuchi, I., Suzuki, T., Kanamori, T., Hachiya, H., Okanohara, D.: Conditional density estimation via least-squares density ratio estimation. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 9, pp. 781–788. Chia Laguna Resort, Sardinia, Italy (13–15 May 2010)

A Semantic Search System for the Supremo Tribunal de Justiça

Rui Melo*

Instituto Superior Técnico, Universidade de Lisboa
rui.melo@tecnico.ulisboa.pt

Abstract. This work developed a prototype of a Semantic Search System to assist the Supremo Tribunal de Justiça (Portuguese Supreme Court of Justice) in its decision-making process. We built a Hybrid Search System that incorporates both lexical and semantic techniques by combining the capabilities of BM25 and the potential of Legal-BERTimbau. In this context, we obtained a 335% increase on the Discovery metric when compared to BM25 for the first query result. This work also introduces a new technique of Metadata Knowledge Distillation.

Keywords: Legal Information Retrieval · Semantic Search · Large Language Models · BERT

1 Introduction

Information retrieval systems evolved significantly over the years. In particular, they can now incorporate advanced techniques from Natural Language Processing (NLP), namely larger and more powerful Large Language Models (LLM) models [1, 6], all of which are based on the groundbreaking Transformers architecture [7]. Information retrieval initially started by utilizing lexical approaches such as the probabilistic model Okapi BM25 (BM25) [3] and has been evolving to incorporate semantic search into the mix. A reliable legal search system is essential for a court to enhance the efficiency and effectiveness of the decision-making process. The legal system is built upon the principle of precedent, which means that previous court decisions serve as a guide for future cases. A reliable search system would allow judges to quickly and accurately access precedents, allow for an extensive and comprehensive coverage of jurisprudence, promoting consistency and transparency. Improving this process would lead to more efficient and informed decisions that uphold the principles of fairness and justice.

2 Semantic Search System Architecture

Our solution to implement a reliable search system involved employing a Bi-Encoder to create independent embeddings for each document. Each sentence

* Master's Program: Bologna Master Degree in Computer Science and Engineering - Taguspark, Instituto Superior Técnico, Universidade de Lisboa
Supervisors: Pedro A. Santos (Instituto Superior Técnico, Universidade de Lisboa), João Dias (Faculdade de Ciências e Tecnologia, Universidade do Algarve)

embedding is generated from the numerous documents using our model, Legal-BERTimbau. Legal documents contain specific language not easily found in conventional websites or books. We collected data from *www.dgsi.pt*, which consists of publicly available STJ court rulings, and indexed them with ElasticSearch[1]. The dataset was divided into three subsets: a training set of 26952 documents, a testing set of 3169 documents, and a validation set of 3169 documents. The text was cleaned and split into singular sentences.

For retrieving specific query results, that exact query would be transformed into an embedding by Legal-BERTimbau. Then the system can proceed to search similar sentences by using the scores provided by BM25 and the cosine similarity value from the embedding space. BM25 scores are normalised. The solution architecture is illustrated in Figure 1.

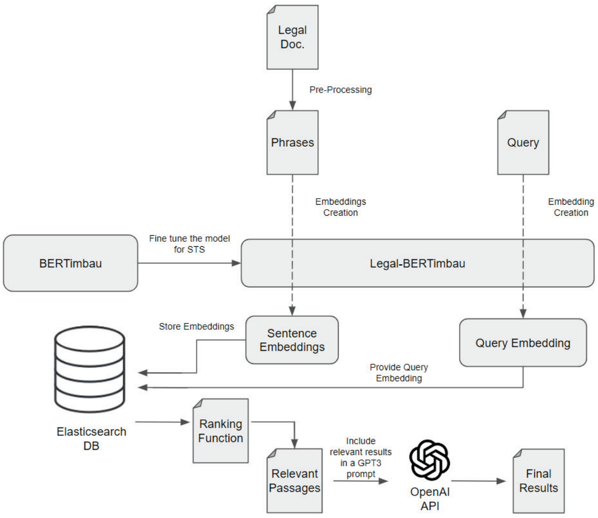

Fig. 1: System Architecture

3 Legal Language Model

We introduce Legal-BERTimbau, a language model adapted to the Portuguese legal domain. We used a foundation BERT model called BERTimbau [5, 4]. We started by adapting the BERTimbau large to the Portuguese legal domain using Masked Language Modeling and the Transformer-based Sequential Denoising Auto-Encoder [8] technique. Furthermore, we fine-tuned the model for Semantic Textual Similarity and explored others techniques with the end goal of improving a bi-encoder's performance. We proceeded to cover Natural Language Inference, Generative Pseudo Labeling [9], Multilingual Knowledge Distillation [2] and even introduced a new technique, Metadata Knowledge Distillation. Experts manually annotate brief tags to identify the main subjects of each document. Documents

[1] https://www.elastic.co/

can have multiple tags, indicating some level of relationship between them. To capture this relationship, centroids are calculated for each tag, and sentence embeddings are adjusted to be closer to their respective centroids. These adjusted embeddings serve as gold labels for training the model.

4 Search System Evaluation

To evaluate the performance of our information retrieval (IR) system, we lacked a pre-existing set of queries and expected results. As a solution, we created embeddings from a collection of 1000 legal documents and generated queries from document summaries using a T5 model. We compared the system's performance by using BM25 searches with the same queries and other multilingual models. The metrics we employed for evaluation were Search and Discovery concepts. The Search metric measures how well a system can find the correct document for a given query. The goal is for the retrieved document to match the one used to create the query. If the retrieved document is the same as the query's source document, the evaluation score increases by 1. The Discovery metric measures how well a search system retrieves relevant documents. Legal documents have one or more manually annotated tags. The Discovery metric increases the score when the retrieved document contains tags that match the tags of the query's original document. For each matching tag, the score is increased by one.

	Search Metric			Discovery Metric		
Model	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3
BM25	**629**	**696**	**722**	685	933	1133
Our System	**629**	675	705	**2984**	**3732**	**4292**
Improvement	0%	-3.01 %	-2.35%	335%	300 %	278%

Table 1: Search System Evaluation

5 Conclusion

This work presents a Search System to enhance information retrieval for legal documents in Portugal. It achievse better results than traditional methods like BM25. This study led to the development of tailored BERT models for our domain, outperforming state-of-the-art multilingual models on specific annotated datasets. The system's performance showed that the proposed search system outperforms the traditional lexical techniques in terms of suggesting similar documents. The implementation of such a search system could drastically improve the information retrieval process and promote consistency in the application of the law.

Acknowledgements

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UIDB/50021/2020

References

1. OpenAI: GPT-4 technical report (2023)
2. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. ACL (2020)
3. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr. **3**(4), 333–389 (apr 2009), https://doi.org/10.1561/1500000019
4. Souza, F., Nogueira, R., Lotufo, R.: Portuguese named entity recognition using BERT-CRF. arXiv preprint arXiv:1909.10649 (2019)
5. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: Cerri, R., Prati, R.C. (eds.) Intelligent Systems. pp. 403–417. Springer International Publishing, Cham (2020)
6. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: LLaMA: Open and efficient foundation language models (2023)
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
8. Wang, K., Reimers, N., Gurevych, I.: TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In: Findings of the ACL: EMNLP 2021. pp. 671–688. ACL, Punta Cana, Dominican Republic (Nov 2021)
9. Wang, K., Thakur, N., Reimers, N., Gurevych, I.: GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In: North American Chapter of the ACL (2021)

Responsible AI, Art, and Finance

Addressing Self-Sustainability in Multi-Agent Systems: Combating Terrorism Financing

David Makiya[1] *[0000−0001−9092−6197]

[1]Department of Informatics, Faculdade de Ciências,
Universidade de Lisboa, Portugal
dnmakiya@fc.ul.pt

Abstract. Self-sustainability of Multiagent Systems (MAS) can be explored by advancing the idea of self-organization (SO) in MAS. We argue that when developing MAS, it must be ensured that they exhibit self-sustainability (SS) as an essential property. The preliminary work presents a consolidated definition of the concept and a set of conditions and properties for a Perfectly Self-Sustainable MAS. Focusing on a specific use-case, combating terrorism financing (TF) within a region, we propose a simple evaluation approach to determine various levels of SS in MAS. While the immediate goal of the concept is to train sustainable agents and to define best practices and conditions on how to achieve this, it anticipates that further down the road, the proposed evaluation methodology of assessing SS within MAS can be used to analyze current practices across other domains and help build a sustainable future. It is a trigger to MAS architects to pay the much needed attention in building systems that are by definition SS either from the application level or within an MAS ecosystem.

Keywords: Multi-agent systems · Sustainability · Terrorism Financing.

1 Introduction

Terrorism is a huge threat to the socioeconomic welfare of many societies. It is safe to assume that terror activities whether state-sponsored or individual have to be financed in one way or another. Starving the perpetrators of such activities from access to their resources forms a solid foundation towards addressing and combating the terrorism threat. Challenges of establishing who a terrorist is, flagging one, tracing their financial patterns, sources of funding and other resources are important in trying to address the issue [7]. The infrequency and small dollar amounts of some funding designs and the indirect relationship between nations and operatives remain the biggest challenge for financial institutions to detect this activity proactively [14]. The patterns being deployed

* PhD Advisor: Prof. João Balsa da Silva[1]
PhD Informatics - Data and Systems Intelligence
Department of Informatics, Faculdade de Ciências, Universidade de Lisboa, Portugal.

by terror groups keep evolving with time. For instance, the National Terrorism Financing Risk Assessment report of the USA identifies that ISIS financial facilitators are constantly looking for ways to consolidate and move funds raised in the US to shell companies in the world hence creating layers within the systems thus giving a vulnerability for transactions to circumvent monitoring by Financial Institutions.

Due to the existence of several layers during the life-cycle of a transaction and complex beneficial ownership structures as hinted in [8, 9], this study proposes to leverage on Multiagent systems to detect terror related funding within banking systems. The drive is to adopt mechanisms that are inherently intelligent to sustainably breakdown a transaction, by delayering it into simple components that can easily be allocated risk weights. The proposed study intends to define the concepts of SS in MAS and adopt them to provide reliable techniques within banking systems to detect as well as prevent terrorism financing. The proposed research will capitalize on exploring SO in MAS and build on this subject to achieve SS [10]. The research will also explore on the approach proposed in [1] on the staggered approach for delayering transactions into stages when building a risk profile for ease of detection [2].

An overview of works shows that sustainability presents itself as a dimension that common solutions fail to address within a MAS environment [4, 13]. Considering that what should be achieved through SS is the ability of the MAS ecosystem to sustain itself without external support within a given time period, then an attempt could be made to assess the best possibility of achieving SS in one of two ways or both. Firstly, within the MAS structured build up Ecosystem and, secondly, from an application level. In the first part, the questionable argument to explore is whether there is need or if it is possible to build a universally acceptable MAS Ecosystem that achieves SS as a standard within its own structure. On the latter, the alternative to the first scenario can be assessed, such that, if it is impossible, then can SS be achieved through specific applications of MAS, dependent on the industry specific problem.

The definitions that SO and SS present should be complementary, however, a preliminary arguable definition of SS can be presented into context as the combined need to have a MAS ecosystem function without a time-bound. With SO in [4, 10], the re-organization of a MAS to its environment is of major concern particularly so that it can deliver the required results by functionality. With SO alone, it seems possible to have a system that can organize/re-organize itself to perform its function and immediately cease to exist. When the element of sustainability is merged with SO it may be possible to realize the essence of having SS within an MAS ecosystem as a preliminary definition of the concept.

2 Research Questions

1. What are the properties of a Self-Sustainable Multi-Agent system/How do the properties of SS-MAS relate with each other?
2. How can machine learning algorithms leveraging MAS be crafted to increase detectability and traceability of TF?

3. Can the National Risk of TF at a point in time be predicted accurately?
4. What quantitative risk metrics should be leveraged from a transaction and/or client to classify it appropriately?

3 Methodology and Model Definition

This section briefly describes the methodology for building the use case setting for combating TF based on the key identified properties of a SS-MAS. Lastly, it proposes an evaluation approach to assess the performance and effectiveness of the MAS.

3.1 Logical Foundation of SS in MAS

Based on the reviews in [3, 5, 6, 11, 12], it becomes apparent that SS is a key property of MAS. The postulation is that SS is a function of scalability (S), flexibility (F), elasticity (E),time (T), purpose (P), social values (V), "Awareness" as (A) and Self-Organization (SO).

Theorem 1. *For a Multi-agent system to be Self Sustainable then, it must pass and satisfy to a TRUE status for S, F, E, T, P, V, A and SO. That is to say, learning and interactions within the MAS ecosystem has to, by* ***necessity*** *and* ***sufficiency****, prove it contains all or a significant level of the conditions at every level (global/macro and local/inner/micro levels). Ideally, if a multi-agent system passes to a TRUE status for both global and micro-levels with the stated conditions, then we can label it a* ****Perfectly Self-Sustainable**** *Multi-agent System.*

3.2 Evaluation Approach

The proposed study shall seek to enunciate the variation in the possible combinations that deviate from the **Perfectly Self-Sustainable** Multi-agent system within all levels. The perfect conditions will be relaxed in an iterative manner for the proposed setting that is described in section 3.3. The study proposes to experimentally evaluate the accuracy and functional adequacy of the proposed setting by relaxing different combinations of conditions and properties to assess which conditions are absolutely necessary or just sufficient to pass the test.

3.3 Proposed Setting

The MAS is a combination of agents aimed at detecting and preventing terrorism financing. The system receives banking data and identifies risk patterns based on client profiles and transactional characteristics. It operates by breaking the detection process into three stages P,Q,R. Each stage is a global Agent class consisting of micro-agents that have specific goals collectively aimed, through co-operation, to achieve the goal of the global agent they belong to. i.e.:

- Predict the National Risk level (P).
- Client Profiling and Transaction Processing (Q)
- Risk Analysis (R)

References

1. Alexandre, C.R., Balsa, J.: A multiagent based approach to money laundering detection and prevention implement of intelligent agents to preventing and combating money laundering crime view project a multiagent based approach to money laundering detection and prevention (2015). https://doi.org/10.13140/2.1.2227.2327, https://www.researchgate.net/publication/271200846
2. Alexandre, C.R., Balsa, J.: Incorporating machine learning and a risk-based strategy in an anti-money laundering multiagent system. Expert Systems with Applications **217**, 119500 (5 2023). https://doi.org/10.1016/J.ESWA.2023.119500
3. Ghadimi, P., Toosi, F.G., Heavey, C.: A multi-agent systems approach for sustainable supplier selection and order allocation in a partnership supply chain. In: European Journal of Operational Research (March 2017)
4. Hoen, P.J.T., Tuyls, K., Panait, L., Luke, S., Poutré, J.A.L.: An overview of cooperative and competitive multiagent learning (2005)
5. Hsieh, F.S.: Scheduling sustainable supply chains based on multi-agent systems and workflow models. In: 2015 International Conference on Intelligent Systems and Knowledge Engineering (April 2015)
6. Ilhan, B., Kog, F.: Bim and sustainability integration:. In: EBF 2019: Advances in Building Information Modeling (2020)
7. de Jesús Rocha-Salazar, J., Segovia-Vargas, M.J., del Mar Camacho-Miñano, M.: Money laundering and terrorism financing detection using neural networks and an abnormality indicator. Expert Systems with Applications **169** (5 2021). https://doi.org/10.1016/j.eswa.2020.114470
8. de Jesús Rocha-Salazar, J., Segovia-Vargas, M.J., del Mar Camacho-Miñano, M.: Detection of shell companies in financial institutions using dynamic social network. Expert Systems with Applications **207** (11 2022). https://doi.org/10.1016/j.eswa.2022.117981
9. Koech, C.: A multi-agent based counter terrorism system through anti-money laundering (2016)
10. Pitt, J.: Self-Organising Multi-Agent Systems. WORLD SCIENTIFIC (EUROPE) (2021). https://doi.org/10.1142/q0307, https://www.worldscientific.com/doi/abs/10.1142/q0307
11. Rovatsos, M., Weiss, G., Wolf, M.: An approach to the analysis and design of multiagent systems based on interaction frames. In: AAMAS '02: Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2 (2002)
12. Serugendo, G.D.M., Gleizes, M.P., Karageorgos, A.: Self-organization in multi-agent systems. Knowledge Engineering Review **0**, 1–24 (6 2005). https://doi.org/10.1017/S0269888905000494
13. Tomičić, I., Schatten, M.: Towards an agent based framework for modelling smart self-sustainable systems. Interdisciplinary Description of Complex Systems **13**(1), 57–70 (2015). https://doi.org/10.7906/indecs.13.1.8
14. Treasury, U.: National terrorist financing risk assessment (2022)

Data Leakage Detection and Data Denoising using Causal Mechanisms for Recommender Systems[*]

Margarida Antunes da Costa[†‡][0000−0002−1799−3382]

Departamento de Ciência de Computadores da Faculdade de Ciências da
Universidade do Porto, Porto, Portugal

Abstract. Due to its capacity to filter content precisely for each user and offer a personalized experience, recommendation systems are becoming increasingly popular. Since they primarily rely on machine learning techniques, they have one significant drawback: their analysis relies on statistical connections, which might come from a wide variety of processes. Recommender systems in the real world rely on user behaviour that can be interpreted as causal. Many outstanding issues in recommendation can be solved using causality. Two data pre-processing tasks are the subject of this study. First, I'll expand the method introduced in [3] to examine data leakage in recommender systems. I will use causal discovery specifically to find the factors that leak information. Second, I'll create a framework for denoising data via causal inference. Finally, as a last use case, I'll test our ideas in recommendation systems for medical decision support.

Keywords: First keyword · Second keyword · Another keyword.

1 Background and State-of-the-Art

Causality

Causality is the relationship between two factors with a logical explanation and changes possibility into actuality. When one variable affects another, for example, it means that the outcome of one is significantly influenced by the other: either the cause occurs before the effect or the result is altered when the reason does.

[*] Supported by organization Associação Portuguesa Para a Inteligência Artificial (APPIA)

[†] **Supervisor:** João Nuno Vinagre Marques da Silva affiliate at the Joint Research Centre, European Comission and Departamento de Ciência de Computadores da Faculdade de Ciêncais da Universidade do Porto
Co-supervisor: Inês Dutra affiliate at CINTESIS@RISE, Porto and Departamento de Ciência de Computadores da Faculdade de Ciêncais da Universidade do Porto

[‡] **Doutoramento em Ciência de Computadores**, Departamento de Ciência de Computadores, Faculdade de Ciências da Universidade do Porto

The causal structural model is employed in semi-supervised learning to explain causality. According to [8], a structural causal model is based on a causal mechanism that generates $P(X_i)$ as all its parents and edges in a structural equation.

Causal inference and causal discovery are the two main subfields of causality research. The first places a focus on directly establishing causality from observable data. The latter aims to determine the impact a modification in a specific variable will have on a desired outcome. Focusing on theoretical components and instruments for integrating causality, "cite" methods-causality finds and studies the ways developed for both jobs.

The counterfactual [9] concept is a crucial idea in this field. The counterfactual is a hypothesis that assesses what would have transpired without a situation.

Causal Mechanisms for Recommender Systems

Causal inference can be used to increase the robustness, interpretability, and reliability of the various recommendation systems, according to the source [5]. For example, debiasing, missing data, and noise are all detected and corrected via causal inference. It also enables us to create more fair and comprehensible models.

Data Leakage

Data leakage [7] happens when a variable attribute contains information about a target variable that cannot exist until the target is known. Leakage is an understudied problem, and models learned from leaky data are unrealistically accurate and should be considered. However, evaluating data in real-world scenario will result in significantly reduction performance. This topic was the basis for my master's thesis, "Data Leakage Detection with Anti-Causal Learning". The result is a hypothetical method investigated and validated by [10] when semi-supervised learning works on a dataset or stream where the target variable Y is the effect of the predictor variable X, there is, X *causes* Y - then there is information from the predictor, which is the effect of the target variable, which contradicts the initial causality claim. We hypothesise that the most plausible explanation for this discrepancy is leakage. To test this hypothesis, we artificially assumed the detection of different kinds of leaks in causal datasets and data streams and signalled leakages when semi-supervised learning outperforms supervised learning.

Data De-noising in Recommender Systems

Recommender systems rely heavily on the history of interactions between users and items to model user preferences. However, in the real world, this information can be noisy. For example, we are on a video application, Youtube, and we give the mobile phone to our child to be entertained. The video he will see will differ from the ones we are used to seeing. So, this is noise information for the recommendation system. According to [2] and [1], recommendation performance suffers from noisy data. The concept has emerged as an enabling area of research, and existing studies, such as [11] [4] [6], show that implicit feedback and noisy communication lead to user satisfaction. We investigate and reinforce differences in degrees.

2 Objective

The main goal of our work is to research and design two frameworks that help us process data to increase the reliability and performance of recommender systems. These two frameworks are data leakage detection and data denoising with causal inference.

Our research questions are, therefore:

- **RQ1:** Can we leverage the causal structure of user feedback data to detect target leaking variables in the recommendation dataset?
- **RQ2:** Can causal models be derived from observed user behaviour without intervention, and can these models be used to identify noisy data points?
- **RQ3:** How does applying leakage detection and denoising to a dataset affect a recommendation model's prediction performance?

To answer these questions, I set the following goals:

- **G1:** Apply causality mechanisms to support leak detection for recommended data [3].
- **G2:** See if the techniques developed in [3] can be applied to recommendation systems. If not, create a new data leakage method for your recommender system.
- **G3:** Develop an open-source library in Python for detecting and remediating data leaks in datasets and data streams from recommender systems.
- **G4:** Use previously developed O2 and O3 libraries for health data to support medical decision-making.

I consider different aspects necessary for improving the model's reliability and intended for use in other domains (health, leisure, finance), thus recommender systems by a causal approach. We want to improve usability.

Acknowledgements: This work is supported by Associação Portuguesa Para a Inteligência Artificial (APPIA) through the contribution of scholarships to students enrolled in higher education courses, in order to promote the participation in scientific events in the area of Artificial Intelligence.

References

1. Amatriain, X., Pujol, J., Oliver, N.: I like it... i like it not: Evaluating user ratings noise in recommender systems. vol. 5535, pp. 247–258 (06 2009). https://doi.org/10.1007/978-3-642-02247-0$_2$4

2. Cosley, Lam, K, S., Tsibouklis, J., Albert, Istvan, Konstan, A, J., Riedl: Is seeing believing?: how recommender system interfaces affect users' opinions (04 2003). https://doi.org/10.1145/642611.642713
3. Costa, M.: Data Leakage Detection with anti-causal learning. Master's thesis, Departamento de Ciência de Computasores da Faculdade de Ciências da Universidade do Porto (2022), `https://hdl.handle.net/10216/146453`
4. Dai, J., Yuan, W., Bao, C., Zhang, Z.: Dgnn: Denoising graph neural network for session-based recommendation (2022). https://doi.org/10.1109/DSAA54385.2022.10032399
5. Gao, C., Zheng, Y., Wang, W., Feng, F., He, X., Li, Y.: Causal inference in recommender systems: A survey and future directions (08 2022)
6. han, Z., N.I., C., W.A., A., Ling, S., Raja, M.: Design of confidence-integrated denoising auto-encoder for personalized top-n recommender systems (02 2023). https://doi.org/10.3390/math11030761
7. Kaufman, S., Rosset, S., Perlich, C.: Leakage in data mining: Formulation, detection, and avoidance. vol. 6, pp. 556–563 (01 2011). https://doi.org/10.1145/2020408.2020496
8. Neuberg, L.G.: Causality: Models, reasoning, and inference, by judea pearl, cambridge university press, 2000. Econometric Theory **19**(4), 675–685 (2003). https://doi.org/10.1017/S0266466603004109
9. Pearl, J., Glymour, M., Nicholas, P.: Causal Inference in Statistics: A Primer. John Wiley & Sons, Ltd (2016), chapter 4 - Counterfactuals and Their Applications
10. Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., Mooij, J.: On causal and anticausal learning. arXiv preprint arXiv:1206.6471 (2012), `https://arxiv.org/pdf/1206.6471.pdf`
11. Wang, W., Feng, F., He, X., Nie, L., Chua, T.S.: Denoising implicit feedback for recommendation. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. ACM (mar 2021). https://doi.org/10.1145/3437963.3441800

Can we hold AGI-enabled Robots Morally Responsible for their Actions?

Mubarak Hussain*$^{[0000-0002-3971-265X]}$

Department of Humanities and Social Sciences, Indian Institute of Technology
Dharwad, Karnataka-580011, India
193101002@iitdh.ac.in

Abstract. This paper argues that if the Artificial General Intelligence (AGI) enabled, robot fulfills functionalist conditions of moral agency, i.e., Interactivity, Independence, and Adaptability prescribed by Floridi and Sanders (2004), AGI-enabled robots could be morally responsible for their actions, at least in a minimal sense. There is a debate in academia about whether DeepMind's AlphaGo is an apt subject for praise/blame for "Move 37". I argue that even if AlphaGo fulfills moral agency conditions prescribed by the functionalists, AlphaGo should not be an apt subject for praise/blame for "Move 37". The researchers of AI and futurists have hypothesized that the developed AGI system would perform any cognitive and behavioral tasks similar to humans. The moral responsibility question may be substantial at that time if an AGI-enabled robot kills an innocent person out of its autonomy. It makes no difference if the AGI system is developed through programming or operates on a computer instead of a biological brain. If the AGI-enabled robot can functionally interact with its environment or other moral agents, it can make any decisions autonomously, adjust to changing circumstances and perform morally relevant actions; they may be held morally responsible, at least in a minimal sense.

Keywords: AI · AGI · moral agency · AMA · moral responsibility · causal responsibility · functionalism.

1 What is AGI?

The AGI appears opposite to Narrow AI [5]. AGI is undoubtedly a young field of study and is still in its early stages of growth. According to Goertzel, the AGI community agrees on the following characteristics of AGI "1. General intelligence involves the ability to achieve a variety of goals, and carry out a variety of tasks, in a variety of different contexts and environments. 2. A generally intelligent system should be able to handle problems and situations quite different from

* Doctor of Philosophy (PhD)
Jolly Thomas, Assistant Professor, Department of Humanities and Social Sciences, Indian Institute of Technology Dharwad, Karnataka-580011, India, joliethomas@iitdh.ac.in

those anticipated by its creators.3. A generally intelligent system should be good at generalizing the knowledge it's gained, so as to transfer this knowledge from one problem or context to others." [5].

2 Moral and Causal Responsibility

Suppose you are an IT employee. Today you perform two significant actions. In the first action, you help a blind couple while crossing the road despite having an important meeting to attend in your company. You would be morally praised in this case since helping needy people is morally right. In the second action, you fix a complicated software issue at your office that no one from your company can fix. In this case, you will also be praised, but not morally. The former action contains a moral element, while the latter does not contain a moral element. People acknowledge that you are not only causally responsible for the result but deserve praise for what you achieved. Now let us examine whether we can hold the AGI system morally responsible based on causal attribution. We can make an argument here:

P1: If the AGI-enabled robot is causally responsible for killing X, AlphaGo is an apt subject for blame

P2: The AGI-enabled robot is causally responsible for killing X

C: Therefore, the AGI-enabled robot is an apt subject to blame

If we accept the conclusion, we must adhere to the implausible notion that causal responsibility is a sufficient condition for normative responsibility, particularly for the appropriateness of praise or blame. A storm is causally responsible for the destruction of a house, but we cannot blame it, or it is inappropriate to blame the storm for its destruction. Nevertheless, the killing of X by the AGI-enabled robot and the destruction brought by the storm are not the same. Here, causal attribution might be a necessary condition but may not be a sufficient condition. To hold the AGI-enabled robot morally responsible for killing X, we must bring the notion of Artificial Moral Agency (AMA).

3 Conclusion

The AMA debate is generally based on two opposite views of moral agency, i.e., the Standard and the Functionalist views. According to the Standard view, to become a moral agent, one must fulfil the conditions of rationality, free will or autonomy, and phenomenal consciousness. However, Functionalists like Floridi and Sanders deny consciousness as the condition of moral agency and hold mindless morality. They prescribed the following three conditions of moral agency [2, 4].

1. Interactivity: The entity can interact with its environment

2. Independence: The entity can change itself and can interact with its environment independently

3. Adaptability: The entity can adjust to changing circumstances

DeepMind's AlphaGo can fulfil the conditions mentioned above of moral agency. However, merely fulfilling the conditions do not make AlphaGo a moral agent. Moral standards do not govern the actions of AlphaGo. To become a moral agent, one must perform morally significant actions, which is missing in AlphaGo. However, holding an AGI-enabled robot morally responsible for its actions from the Functionalist perspective may be possible, as functionalism holds that mental states can be understood in terms of what an entity can do rather than how it is made of [12]. Philosophers usually consider two exclusively necessary and jointly sufficient conditions to hold someone morally responsible for their actions [13]. The first condition is a control, and the second is an epistemic. Functionalists may accept the control conditions but deny epistemic conditions. As per the traditional or standard view, to hold someone morally responsible (or to be praised or blamed) for one's action, she must be aware of four probable epistemic requirements while acting, i.e., i) awareness of the action, ii) awareness of the moral significance, iii) awareness of the consequence, and iv) awareness of the alternatives [13]. However, how do we know whether someone has such awareness while acting? This issue directs us to the philosophical problem of other minds. I am sure I have a mind and mental state. I am directly accessible to my mind and mental states. However, I have no direct knowledge of anyone with minds like mine. Therefore, epistemic conditions may be insignificant in holding someone morally responsible. Computational functionalism holds that mental states like pain, beliefs, desire, and thoughts are computational states of the brain [14]. Looking into the characteristics of AGI, we may say that the Functionalist conditions of moral agency, i.e., Interactivity, Independence, and Adaptability, are fulfilled by the AGI-enabled robot and qualify as an AMA. If the AGI-enabled robot may perform morally significant actions along with the three conditions of AMA, it may be an AMA. AGI's intellect system would indeed differ from humans. However, they may still display intelligence and moral behaviour similar to moral agents. It makes no difference if the AGI is developed through programming or operates on a computer instead of a biological brain. In the case of humans, when we say that someone has a moral agency, we also consider the conditions mentioned above of moral agency. The internal state or consciousness condition is not crucial while conferring moral agency since we do not have access to or consider what type of consciousness someone had while performing a specific action. Until and unless someone's intention is functionally manifested in a specific action, we cannot determine the kind of consciousness or awareness they are experiencing while acting. Therefore, if the AGI-enabled robot can functionally interact with its environment or other moral agents, it can make any decisions autonomously, adjust to changing circumstances and perform morally relevant actions; they may be held morally responsible, at least in a minimal sense.

Acknowledgement

This work was supported by TiHAN Project (EE/AM/TiHAN/2021/8.57).

References

1. Allen, C., Varner, G., Zinser, J.: Prolegomena to any future artificial moral agent. Journal of Experimental & Theoretical Artificial Intelligence **12**(3), 251–261 (2000). https://doi.org/10.1080/09528130050111428
2. Behdadi, D., Munthe, C.: A normative approach to artificial moral agency. Minds and Machines **30**(2), 195–218 (2020). https://doi.org/10.1007/s11023-020-09525-8
3. Bostrom, N., Yudkowsky, E.: The ethics of artificial intelligence. The Cambridge handbook of artificial intelligence. Cambridge University Press (2014)
4. Floridi, L., Sanders, J.W.: On the morality of artificial agents. Minds and machines **14**, 349–379 (2004). https://doi.org/10.1023/B:MIND.0000035461.63578.9d
5. Goertzel, B.: Artificial general intelligence: concept, state of the art, and future prospects. Journal of Artificial General Intelligence **5**(1), 1–48 (2014). https://doi.org/10.2478/jagi-2014-0001
6. GoogleDeepMind: Alphago. https://www.deepmind.com/research/highlighted-research/alphago (2023), last accessed 18 May 2023
7. Hakli, R., Mäkelä, P.: Moral responsibility of robots and hybrid agents. The Monist **102**(2), 259–275 (2019). https://doi.org/10.1093/monist/onz009
8. Johnson, D.G.: Computer systems: Moral entities but not moral agents. Ethics and information technology **8**, 195–204 (2006). https://doi.org/10.1007/s10676-006-9111-5
9. Levin, J.: Functionalism. https://plato.stanford.edu/entries/functionalism/ (2023), last accessed 12 May 2023
10. Mele, A.: Moral responsibility for actions: Epistemic and freedom conditions. Philosophical Explorations **13**(2), 101–111 (2010). https://doi.org/10.1080/13869790903494556
11. Pennachin, C., Goertzel, B.: Contemporary approaches to artificial general intelligence. In: Artificial general intelligence, pp. 1–30. Springer (2007)
12. Polger, T.W.: Functionalism. https://iep.utm.edu/functism/ (2023), last accessed 15 May 2023
13. Rudy-Hiller, F.: The epistemic condition for moral responsibility. https://plato.stanford.edu/entries/moral-responsibility-epistemic/ (2022), last accessed 8 May 2023
14. Shagrir, O.: The rise and fall of computational functionalism. Cambridge University Press (2005)
15. Tigard, D.W.: Artificial moral responsibility: How we can and cannot hold machines responsible. Cambridge Quarterly of Healthcare Ethics **30**(3), 435–447 (2021). https://doi.org/10.1017/S0963180120000985
16. Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press (2009)

Symbolic music generation conditioned on continuous-valued emotions

Serkan Sulun[*]

Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), 4200-465 Porto, Portugal
serkan.sulun@inesctec.pt

Abstract. In this paper we present a new approach for the generation of multi-instrument symbolic music driven by musical emotion. The principal novelty of our approach is the conditioning of a state-of-the-art transformer based on continuous-valued valence and arousal labels. In addition, we provide a new large-scale dataset of symbolic music paired with emotion labels in terms of valence and arousal.

Keywords: music generation · MIDI · transformers · emotion · affective computing

1 Introduction

In this work, we introduce an affective algorithmic composition model that can be conditioned on continuous coordinates on the valence-arousal plane. To this end, we combine several datasets, resulting in a labeled MIDI dataset two orders of magnitude greater than existing labeled MIDI datasets. Using this dataset, we successfully train large transformer models [17] on a single GPU, to generate multi-instrument symbolic music conditioned on emotion. To the best of our knowledge, this is the first music generation model that can be conditioned on both valence and arousal simultaneously, therefore enabling conditioning on an arbitrary emotion from the widely-used circumplex model of affect.

2 Related work

The creators of the VGMIDI dataset devised a method for symbolic music generation conditioned on emotion [4]. Using a genetic algorithm, they fine-tuned the weights of a pretrained LSTM. This was done separately for positive and negative valence conditions, resulting in two models. Both Hung et al. and Zhao et al. generated symbolic music conditioned on four categorical emotions belonging

[*] Doctoral Program in Electrical and Computer Engineering at the University of Porto. Supervisors: Paula Viana, Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Porto, Portugal and ISEP, Polytechnic of Porto, School of Engineering, Porto, Portugal

to the four quadrants of the valence-arousal plane [7, 19]. Zhao et al. [19] labeled the piano-midi dataset [10] using categorical labels, and trained a biaxial LSTM [8] on this labeled dataset. Hung et al. [7], the creators of the EMOPIA dataset, trained a transformer model that is conditioned using control tokens [9].

The *Spotify for Developers* API allows users to access audio features for a given song from Spotify's private database [14]. These audio features are both low- and high-level and are namely danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, and tempo. The high-level features such as valence are estimated using machine learning algorithms that are trained on data labeled by experts [13, 16].

3 Methodology

3.1 Lakh-Spotify dataset

To create a dataset that contains pairs of MIDI files and high-level labels, we use the Spotify for Developers API and obtain audio features for the corresponding samples from the Lakh MIDI dataset (LMD) [12]. In particular, we use the *LMD-matched* subset, since its samples are matched to the entries in the Million Song Dataset (MSD) [2], hence we can use the metadata from MSD to search Spotify's database. Using the track ID for each MIDI file, we first obtain the song title, artist name, and Echo Nest song ID. Using the Echo Nest song IDs, and another dataset named *Million Song Dataset Echo Nest mapping archive* [1], we also obtained Spotify track IDs.

Next, for each MIDI sample, we conducted a search using the Spotify for Developers API to obtain audio features. While the precise implementation details for their retrieval are not publicly available, an explanation of the Spotify audio features can be found in the online documentation[1]. The query for the search was the associated Spotify ID. If the Spotify ID was not available, we used the artist name and the song title as the query. We also extracted and included the low-level MIDI features such as note density, i.e., number of notes per second, the estimated tempo, and the number of instruments. These low-level features can also be used to model the arousal dimension of the circumplex model of affect [15, 18].

3.2 Emotion-based music generation

Training data and pre-processing For our music generation task, we first pre-train our non-conditional vanilla model on the Lakh Pianoroll Dataset (LPD) [3], specifically the LPD-5-full subset. This subset is created by merging the individual tracks in the MIDI files into five common categories: drums; piano; guitar; bass; and strings. After pre-processing, the non-conditional training data split has 96119 songs.

[1] `https://developer.spotify.com/documentation/web-api/reference/`

To train our conditional models, we first transfer the available weights from the vanilla model and then fine-tune on the LPD-5-matched dataset, namely the 5-instrument piano roll counterpart of the LMD-matched dataset. Since we previously generated low- and high-level labels for this dataset as explained in Section 3.1, we used these labels for conditioning. After pre-processing, the conditional training data split has 27361 songs. We transpose the pitches of all instruments except drums, by a randomly chosen integer value between -3 and 3 inclusive. We use two conditioning values to model valence and arousal, the valence feature from Spotify's database and the MIDI note density averaged over the number instruments respectively.

Models The backbone of our models is the music transformer [6], which is a decoder-only transformer using relative position embeddings. It has 20 layers and a feature dimension of 768. Each layer has 16 heads and a feed-forward layer with a dimension of 3072. Overall, our model has around 145 million parameters.

We experimented with different methods for conditioning the music generation process on the emotion features. We first implemented the current state-of-the-art approach in conditional sequence generation [7, 9, 11], which we name *discrete-token*, where we put the valence and arousal values into discrete bins and then converted them into control tokens. In detail, we quantize the condition values using 5 equal-sized bins, where the central bin index is 0. The control tokens belonging to valence and arousal are placed before the music tokens, i.e., concatenated in the sequence dimension. In our next approach, named *continuous-token*, we use the normalized condition values in their continuous form. We feed each value to a separate linear layer, creating condition vectors that have the same length as the music token embeddings. Next, the condition vectors and music token embeddings are concatenated in the sequence dimension and fed into the transformer. Our final approach is named *continuous-concatenated*, where we create a single vector for the two normalized continuous condition values, repeat this vector in the sequence dimension, and concatenate it with every music token embedding. The lengths of the conditioning vectors and token embeddings are 192 and 576 respectively, so that the total feature length of the transformer input is constant across models.

At inference, and before the generation starts, the input sequence only consists of the `<START>` token, except for the *discrete-token* model where we also prepend two condition tokens for valence and arousal. For the models *continuous-token* and *continuous-concatenated*, the condition values are fed in parallel at every timestep. We generate the output autoregressively, where the generated token is appended to the input sequence, forming the input sequence for the next timestep. We use nucleus sampling with $p = 0.7$ from the temperature adjusted softmax distribution [5], with a temperature of 1.2. To avoid excessive repetitions, if the number of tokens in the nucleus in the previous step was less than 3, we increase the temperature slightly. The code, trained models, qualitative output samples, and extended paper with quantitative results are available online[2].

[2] `serkansulun.com/midi`

References

1. AcousticBrainz: Million song dataset echo nest mapping archive, https://labs.acousticbrainz.org/million-song-dataset-echonest-archive/, accessed: 2023-06-13
2. Bertin-Mahieux, T., Ellis, D.P., Whitman, B., Lamere, P.: The million song dataset. In: Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011) (2011)
3. Dong, H.W., Hsiao, W.Y., Yang, L.C., Yang, Y.H.: Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In: Thirty-Second AAAI Conference on Artificial Intelligence. pp. 34–41 (2018)
4. Ferreira, L., Whitehead, J.: Learning to generate music with sentiment. In: Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, the Netherlands, November 4-8, 2019. pp. 384–390 (2019)
5. Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The curious case of neural text degeneration. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020 (2020)
6. Huang, C.Z.A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., Dai, A.M., Hoffman, M.D., Dinculescu, M., Eck, D.: Music transformer: Generating music with long-term structure. In: International Conference on Learning Representations (2018)
7. Hung, H.T., Ching, J., Doh, S., Kim, N., Nam, J., Yang, Y.H.: Emopia: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. In: Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021. pp. 318–325 (2021)
8. Johnson, D.D.: Generating polyphonic music using tied parallel networks. In: Correia, J., Ciesielski, V., Liapis, A. (eds.) Computational Intelligence in Music, Sound, Art and Design. Lecture Notes in Computer Science, vol. 10198, pp. 128–143. Springer (2017)
9. Keskar, N.S., McCann, B., Varshney, L.R., Xiong, C., Socher, R.: Ctrl: A conditional transformer language model for controllable generation. arXiv preprint arXiv:1909.05858 (2019)
10. Krueger, B.: Classical piano midi page, http://www.piano-midi.de, accessed: 2023-06-13
11. Payne, Christine: Musenet (Apr 2019), https://openai.com/blog/musenet, accessed: 2023-06-13
12. Raffel, C.: Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-Midi Alignment and Matching. Ph.D. dissertation, Dept. Graduate School Arts Sci., Columbia Univ., New York, NY, USA (2016)
13. Skidén, P.: New endpoints: Audio features, recommendations and user taste, https://web.archive.org/web/20221211110212/https://developer.spotify.com/community/news/2016/03/29/audio-features-recommendations-user-taste/, accessed: 2023-06-13
14. Spotify: Spotify for developers, https://developer.spotify.com, accessed: 2023-06-13
15. Tan, H.H., Herremans, D.: Music fadernets: Controllable music generation based on high-level features via low-level feature modelling. In: Proceedings of the 21st International Society for Music Information Retrieval Conference, ISMIR 2020 (Oct 2020)

16. The Echo Nest: Plotting music's emotional valence, 1950-2013, `https://web.archive.org/web/20180604003930/https://blog.echonest.com/post/66097438564/plotting-musics-emotional-valence-1950-2013`, accessed: 2023-06-13
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
18. Williams, D., Kirke, A., Miranda, E.R., Roesch, E., Daly, I., Nasuto, S.: Investigating affect in algorithmic composition systems. Psychology of Music **43**(6), 831–854 (2015)
19. Zhao, K., Li, S., Cai, J., Wang, H., Wang, J.: An emotional symbolic music generation system based on lstm networks. In: 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). pp. 2039–2043 (2019)

Exploring GenAI Art Tools By Designing Visual Culture Chatbot Assistant: A Case Study In Higher Education

Yi-yang Liu[1][0000−0002−9662−1500]*

Faculty of Engineering, Free University of Bolzano-Bozen,39100, Italy
yiyang.liu@stud-inf.unibz.it

Exploring GenAI Art Tools By Designing Visual Culture Chatbot Assistant

Abstract. The capability of generative AI (GenAI) art tools nowadays alters the creativity process due to the advantage of the deep learning architecture of transformers and Large Language Models(LLMs). However, the style-oriented focus of prompt-based AI art tools might cause us to ignore the metaphoric connections behind symbols, which not only leads to an educational crisis in art but also in the algorithms. According to EMT (Extended Mind Thesis), both human and non-human entities could gain intelligence that fits the new paradigm of Hybrid intelligence (HI). Therefore, this study proposes to design a chatbot assistant with customized domain knowledge of visual culture art education(VCAE) and a persona trained by dialogic pedagogy to challenge the mainstream GenAI art tools, the hypothesis is to fill the knowledge gaps by offering deeper thoughts during the human-machine interaction. A case study will be conducted together with University lecturers and students in higher education mastered in art and design, and serve as an iterative process for improving the application. The objective is to discover humanity in GenAI art tools influenced by VCAE.

Keywords: generative-AI · Visual culture art education · Hybrid intelligence · AI chatbot · Extended Mind Thesis.

1 Introduction

1.1 AI meets HCI=HI

The development of chatbots has made significant progress in recent years. Due to the enormous success in the development of large language models (LLM), conversational AI has gained significant attention and demonstrated promising results in various industries and scenarios (Nagarhalli et al., 2020), such as parent training (Entenberg et al., 2021) and promoting students' "trust, self-confidence,

* Doctoral Program in Computer Science
Advisors:
Antonella De Angeli[1] (antonella.deangeli@unibz.it)

and open-mindedness" (Muthmainnah et al., 2022). In the heavily disrupted education industry, this means that conversations with AI chatbots like ChatGPT can inspire individuals to engage in thought-provoking discussions, exploration, and even innovative ideas. For example, enhancing critical thinking in STEM learning (Vasconcelos et al., 2023). However, concerns such as morality, bias, and correctness also require a shift in pedagogy and a rethinking of traditional intelligence in the new era. (Taecharungroj, 2023). Consequently, there is a growing interest in HCI to explore the new paradigm of the Human–AI collaboration process based on concepts such as hybrid intelligence or an earlier term called Extended Intelligence, both of which acknowledge the concept of intelligence as "a fundamentally distributed phenomenon" (Ito, 2016). Inspired by the lens consider that AI is more than just building better machines; it is a way "to use machines to understand the mind itself" (Minsky, 2007). Despite the enormous impact, the focus of developing chatbots is mostly on how to understand humans better and generate more satisfactory outcomes. With the increasing interest in the mutual learning HI concept, various authors indicated that this might have overshadowed the potential risks of AI lacking the ability to provide deeper insights and reflections other than the skills and knowledge obtained (Bredeweg et al.; Chen et al.,2022; Gillani et al.; Roumate, 2023). To address this gap, the proposed doctoral study will critically review the development and limitations of how the current experience of using GenAI art tools is insufficient, then foster critical dialogue and deepen understanding by developing an app trained with the domain knowledge of visual culture in order to make a contribution to the future HI paradigm.

1.2 Research Questions

RQ1 - Pilot study phase: **What insight could visual culture arts educators and learners provide for the experience of prompt-based interactive creativity using Gen AI's art creation tools?** -Addressed by a literature review, anthropological observation, and in-depth interviews with experts in the field of VCAE. RQ2 - Chatbot design phase: **How can domain knowledge in dialogic theory and visual culture art education be structured in the LangChain framework for developing a new paradigm of HI in art creation?** -With the intention of reflecting on human creativity in a prompt-based GenAI tool through VCAE domain knowledge and the dialogical pedagogy, which will be applied to the LangChain framework to enable the text-to-image/ image-to-text AI chatbot assistant with a more engaging experience. RQ3 - User experience evaluation phase: **How can the user experience of the text-to-image AI chatbot assistant be evaluated in order to meet the requirements of Hybrid Intelligence?** RQ3-1. **From the perspective of Hybrid Intelligence, what is the impact of the creative experience of the customized AI visual culture chatbot assistant on the user's critical aesthetic thinking?** RQ3-2. **From the perspective of Human in the loop service design, how does this study of customized AI vi-**

sual culture chatbots respond to the concerns of a responsible and explainable AI?

2 State of the art

2.1 Mainstream GenAI art tools

Applications like Stable Diffusion, Midjourney, and Dall-E 2 have made significant advancements in the field of generative models, particularly in generating and editing high-quality images. The main criticism of the text-to-image, prompt-based procedure is that it can sometimes produce unexpected or undesirable results due to the limitations of the generative models in understanding the context and nuances of the textual prompts. This can lead to images that may not accurately represent the intended concept or may contain inappropriate content. Some tools also offer image-to-image editing, where users can upload an existing image and modify it based on text prompts, further expanding the possibilities of co-creation with AI in the visual arts domain. Regarding the image-to-text function, Midjourney and other models like OpenAI's CLIP can perform tasks reversely, this is extremely valuable since it offers humans an elementary way to learn from how AI interprets images and gets inspired.

2.2 LangChain framework and LLMs API

It was until late 2022, the Interest in LLMs and generative AI brought OpenAI's ChatGPT to skyrocket in popularity. However, the real potential of LLMs which represents a notable advancement in the evolution of chatbots lies in the integration of specialized knowledge and the openness protocol. LangChain, a framework built around LLMs, exemplifies this progress by offering Python libraries that streamline the development process for language model-powered applications. By providing an array of tools, components, and interfaces, the framework not only enhances the functionality and effectiveness of chatbots but also simplifies their overall development process, which provides a concrete reference for the app development planned in this study.

3 Methodology

In their 2007 study, Harrison et al. categorized the methodological frameworks of HCI into three distinct paradigms. Each of these paradigms represents a different worldview and contributes to HCI in different ways. This research incorporates the third paradigm called the Phenomenologically-Situated Paradigm, which assumes researchers "should not just react to the changes around it but also shape those changes" (Dix, 2017), and proposes a shift toward designing for the broader context of social and cultural influences.

References

1. Nagarhalli, T. P., Vaze, V., Rana, N. K. A Review of Current Trends in the Development of Chatbot Systems. 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 706–710. (2020)https://doi.org/10.1109/ICACCS48705.2020.9074420
2. Entenberg, G. A., Areas, M., Roussos, A. J., Maglio, A. L., Thrall, J., Escoredo, M., and Bunge, E. L. Using an Artificial Intelligence Based Chatbot to Provide Parent Training: Results from a Feasibility Study. Social Sciences, 10(11), 426. (2021)https://doi.org/10.3390/socsci10110426
3. Muthmainnah, Ibna Seraj, P. M., Oteir, I. Playing with AI to investigate human-computer interaction technology and improving critical thinking skills to pursue 21st century age. Education Research International, 2022, pp. 1–17. (2022)https://doi.org/10.1155/2022/6468995
4. Vasconcelos, M. A. R., dos Santos, R. P. Enhancing STEM Learning with ChatGPT and Bing Chat as Objects to Think With: A Case Study. (2023) http://arxiv.org/abs/2305.02202
5. Taecharungroj, V. "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. Big Data and Cognitive Computing, 7(1), 35. (2023) https://doi.org/10.3390/bdcc7010035
6. Ito, J. Extended Intelligence. Joi Ito's PubPub. (2016)https://doi.org/10.21428/f875537b
7. Minsky, M. The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind. Simon and Schuster. (2007)
8. Bredeweg, B., Kragten, M. Requirements and challenges for hybrid intelligence: A case study in education. Frontiers in Artificial Intelligence. (2022) https://doi.org/10.3389/frai.2022.891630
9. Chen, X., Zou, D., Xie, H., Cheng, G., Liu, C. Two Decades of Artificial Intelligence in Education: Contributors, Collaborations, Research Topics, Challenges, and Future Directions. Educational Technology and Society, 25(1), 28–47. (2022)https://www.jstor.org/stable/48647028
10. Gillani, N., Eynon, R., Chiabaut, C., Finkel, K. Unpacking the "Black Box" of AI in Education. Educational Technology and Society, 26(1), 99–111. (2023)https://www.jstor.org/stable/48707970
11. Roumate, F. Artificial Intelligence in Higher Education and Scientific Research: Future Development. Springer Nature. (2023)
12. Dix, A. Human–computer interaction, foundations and new paradigms. Journal of Visual Languages and Computing, 42, 122–134. (2017)https://doi.org/10.1016/j.jvlc.2016.04.001