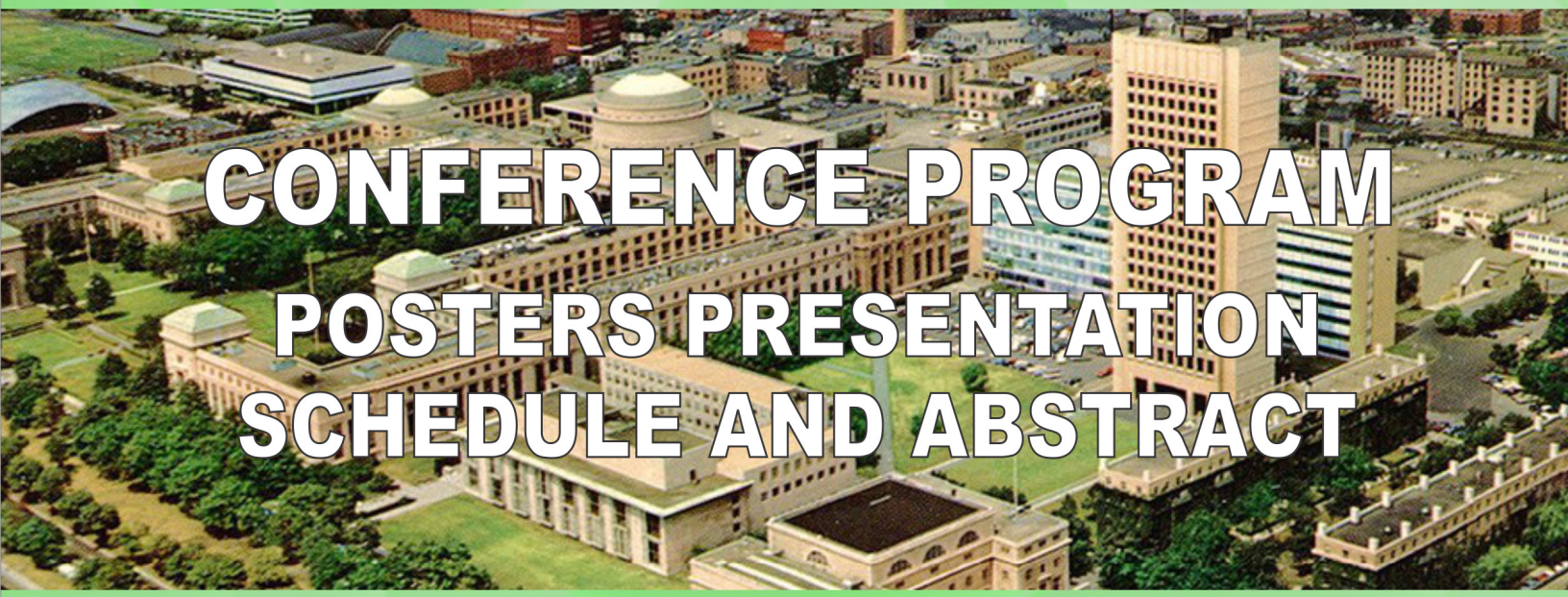
IEEE MIT URTC 2020

UNDERGRADUATE RESEARCH TECHNOLOGY CONFERENCE

October 09 - 11, 2020 | Cambridge, Massachusetts, USA (Virtual)

MEET INNOVATIVE TECHNOLOGY



CONFERENCE PROGRAM
POSTERS PRESENTATION
SCHEDULE AND ABSTRACT

Organized and Sponsored by IEEE Boston Section and MIT IEEE Student Branch

https://urtc.mit.edu

mit ieee
urtc 2020

IEEE Catalog Number: CFP20E50
ISBN: 978-1-7281-7571-3

October 11, 2020 (Sunday)

Technical Poster Presentation (Session #1 Room B)

EST 8:00am - 10:00am

Machine Learning / Artificial Intelligence (AI) Track

Track Chair: Penny Tan

➢ 8:10am (PO20-0012)

Coreference Resolution for Neutralising Gendered Pronouns

Parth Raghav (Reed College)

Gender Neutralization is an important task in text anonymization and generatively producing gender-free descriptions of people and objects. We demonstrate a tool that utilizes Coreference Resolution and a heuristic to neutralize long gendered texts. English is a grammatically gendered language. As gender is socially classified into two exclusive genders, male and female, our language reflects this classification, also known as the Gender Binary, in its syntax. Gendered syntax prevents accurate third-person descriptions of non-binary and transgender identifying folks. Moreover, it reinforces gender-related stereotypes " using masculine pronouns for upscale jobs and feminine pronouns for jobs relating to indoor work. Policeman and stewardess are two popular examples. Primer on Gender-Neutral Language A gender-neutral language on the contrary does not presuppose a particular natural gender and as a result, deals away with all problems accounted for above. LGBTQ activists and linguists have championed more inclusive language by creating new non-binary pronouns and retooling already existing pronouns and grammatical constructions. Despite their adoption in Merriam-Webster Dictionary, Associated Press Stylebook and Chicago Manual of Style, mainstream adoption isn't easy. A popular argument against œthey/them as both singular and plural pronoun is that it muddies the syntax, making it harder to comprehend longer sentences. However, in practice, this generally doesn't hold true, as we will show in this demonstration. Automated text anonymization is a solution for hiding instances of a name from documents. However, this is a challenging task due to the unstructured form of textual data. On top of replacing a name with a pseudonym, the solution must also toggle the grammatical gender. Thus, Gender Neutralization is an essential task in the text anonymization pipeline. To help parents and trans-allies transition to gender-neutral English, we created a chrome extension to neutralize gendered pronouns on all webpages across the browser to augment user's web experience with a gender-free view. All code available at https://github.com/parthraghav/Pronoun-Pro 2.1 Previous work Previous attempts [2,3] to solve the problem relied on context-unaware pattern matching, also known as Regular Expression, and performed poorly since replacing gendered pronouns with non-binary pronouns often obfuscates sentences and reduces comprehension. Our solution comprises of a client-facing content script that transforms the webpage and a server that manipulates the data. The content script crawls the body of the page and parses big chunks of text for entities and sends a request to the server to annotate and resolves coreference clusters in the document. Coreferences are linguistic expressions that refer to the same entity. From an experimental analysis, we derived a heuristic below to maximize comprehension. 1) If a sentence has a collective noun, the pronoun is instead resolved into the main mention. 2) If the distance between the pronoun and its closest coreference is higher than the Attention Threshold*, the pronoun is not neutralized. 3) If a sentence has more than a specific number of pronouns, as dictated by the Obscurity threshold*, the pronoun that is nearest to its coreference is neutralized, the remaining are resolved into their main mentions. Attention threshold refers to the number of sentences program will recursively lookup to find a coreference. Obscurity threshold refers to the number of gender-neutral pronouns that a sentence can have without confusing the reader. The default values of these parameters are computed based on the dependency structure of the current sentence and the paragraph. 2.3 Broader Impact Current political climate affects transgender people the most, causing more hate crimes than ever and ever so increasing suicide rates of Black-trans people in America. Transphobia is disrupting more lives than ever, and yet technology is going unused to solve problems that are faced by the LGBTQ community. There's a need to create web tools for the next generation to leap out of the ignorance that often matures to take shape of transphobia. We posit if we erased the gender-binary from the web, our society will shortly follow. 3. Interactivity & Virtual Environment Our poster includes a demo where participants can fiddle with the algorithm, a live server, and a chrome extension that can be downloaded and installed if participants continue to use the service. Demo Users can enter the URL of their favorite news article or essay that they wish to read or choose any of the suggestions down below the textbox. Demo server uses Google Rendertron to render Javascript-enabled websites and returns the text body of the webpage. Users can choose to either neutralize or anonymize the text. Users can view both the original text and the converted text side-by-side to compare text readability for them. There are two adjustable parameters: (i) Attention threshold, and (ii) Obscurity threshold. Changing each parameter reflects a systematic change in the comprehensibility metric of the text. Users can also toggle annotations for the neutralized text to better understand how each coreference relates to each other. In the spirit of making demo participants comfortable and confident with gender-neutral language, we also made an inline quiz for participants to guess which entity the gender-neutral pronoun refers to. The task for users and servers alike is to maximize the score. Similarly, there would a bug bounty for participants to catch grammatical errors that our model missed. We would use the collected data to make our model better at neutralizing gendered grammar and publish statistics on the demo website.

➢ **8:20am (PO20-0047)**

Faster Region-Based Convolutional Neural Network for Tumor Localization in Thermograms: A Novel Deep-Learning Based Approach to Breast Cancer Screening

Sadhana Lolla (Massachusetts Institute of Technology)

One in eight women in the United States will be diagnosed with breast cancer in her lifetime. Early diagnosis of breast cancer is key to effective treatment and an increased survival rate. Mammography, the current gold standard of diagnosis, exhibits a low sensitivity in younger women and women with dense breast tissue and has a high false positive rate. Alternate forms of screening include breast thermography, which does not require compression of the breast and is equally accurate regardless of age or tissue density. However, breast thermography has a much higher false positive rate and a much lower true positive rate than mammography. Computer Aided Diagnosis (CAD) systems can help reduces the false positive rate and increases the sensitivity of thermography while retaining its benefits over mammography. To date, however, there have been no CAD systems developed to identify tumors in breast thermograms. In this study, we aim to develop a CAD system that can localize tumors in thermograms with a comparable sensitivity and specificity to that of radiologists with mammograms. We develop a Faster Region-Based Convolutional Neural Network with a sensitivity of 0.906, a false positive rate of 0.232, and an AUC of 0.89. This CAD system not only outperforms radiologists, but also is more sensitive than other algorithms designed to detect tumors in mammograms, ultrasound, or other modalities. This paper demonstrates that in conjunction with this CAD model, breast thermography can be used as a reliable method of screening and is a step forward in the field of computer aided diagnostics.

➢ **8:30am (PO20-0049)**

A Novel Pupillometric Based Application for the Automated Detection of ADHD Using Machine Learning and Time-series Analysis

William Das, Shubh Khanna (Hunter College High School, Icahn School of Medicine at Mount Sinai)

Accurate and efficient detection of attention-deficit/hyperactivity disorder (ADHD) is critical to ensure proper treatment for affected individuals. Current clinical examinations, however, are inefficient and prone to misdiagnosis, as they rely on qualitative observations of perceived behavior and are subject to inherent doctor biases. This has perennially resulted in lengthy clinical examinations lasting multiple hours, and a high misdiagnosis rate, estimated in 2010 at 20%. Wait times from initial concern to diagnosis can extend beyond 13 months, as the demand for examinations often exceeds the capacity of pediatric clinics, preventing the administration of proper treatment regimens for affected individuals. We propose a robust machine learning based framework in conjunction with a novel and tangible web application that analyzes pupil-size dynamics as an objective biomarker for the automated detection of ADHD. Our framework integrates a comprehensive pupillometric feature engineering and visualization pipeline with state-of-the-art binary classification algorithms and univariate feature selection. An evaluation of various classifiers showed that the support vector machine achieved an optimal average 85.6% area under the receiver operating characteristic (AUROC), 77.3% sensitivity, and 75.3% specificity using 10-fold nested cross-validation (CV) on a declassified dataset of 50 patients. 218 of the 783 engineered features, including fourier transform metrics, absolute energy, consecutive quantile changes, approximate entropy, peak numbers, aggregated linear trends, local maxima and minima, overall variation, as well as pupil-size dilation velocity, were found to be statistically significant differentiators ($p < .05$), and provide novel behavioral insights into associations between pupil-size dynamics and the presence of ADHD. In particular, ADHD positive subjects exhibited predominantly lessened variation in pupil-size across a time-series, while displaying higher approximate entropy, or unpredictability in pupillary responses. Furthermore, ADHD positive subjects exhibited lessened local maxima, minima, and pupil-size dilation velocity in response to visual stimuli than healthy subjects, indicating decreased neural activity and attentional performance. RadViz multivariate projections were also constructed, and highlight the strong class separability enabled by the engineered features. Despite a limited sample size, the strong AUROC values highlight the robustness of the binary classifiers in detecting ADHD as such, with additional data, sensitivity and specificity metrics can be substantially augmented. The optimal model was incorporated in a tangible web application that administers standard visuospatial working memory task, lasting ~20 minutes, and outputs a probability of diagnosis using eye biometrics captured in real-time. The application employs convolutional neural networks, circle hough transform methods, and a custom ray tracing algorithm to segment the pupil in real-time. Transcending the boundaries of simple diagnostics, our application provides a probabilistic risk score of having ADHD in conjunction with medical advice on whether a patient should consult a medical practitioner or neurologist. This study is the first to apply machine learning based methods for the detection of ADHD using solely pupillometrics, and highlights its strength as a potential discriminative biomarker to drive clinical management and decision-making, paving the path for the development of novel diagnostic applications to aid in the detection of ADHD using oculometric paradigms and machine learning.

➢ **8:40am (PO20-0051)**

Fast Detection of Toxic Emissions Using Deep Neural Network Based Sensing

Ibrahim Bhavnagarwala, Adam Bhavnagarwala (Danbury High School, NJIT)

We demonstrate, using a simple inexpensive IoT system, equipped with an array of gas sensors and WiFi connectivity, the ability of a DNN to quickly identify a toxic gas by recognizing patterns even of small changes in concentrations of its component gases. These patterns are recognized at very low component gas concentrations enabling a DNN based gas array sensor to provide early and accurate detection while toxic emissions still have low concentrations. The speed of DNN based detection is limited only by the speed of gas diffusion to the sensor arrays enabling the toxic gas detection to take place much sooner than conventional sensor-based detectors that rely on the concentration of the component gases exceeding a critical threshold. We also demonstrate the ability of DNNs to classify relative toxicity and flammability of gases commonly found in Hospitals (Anesthetic's, aerosolized medications and chemicals used as a fixative such as formaldehyde, toulene etc.), Waste water treatment plants, Restaurants (CO, CO2, N2, CH4), Mechanical/boiler rooms (refrigerants), Pharmaceutical Labs (HCN), Oil refineries (BTEX), Cold storage (NH3) and Industrial manufacturing. These classifications help Hospitals, Workplaces and Schools to monitor toxicity and flammability according to the tolerance people have in designated areas to cigarette smoke or other toxic gas emissions.

➢ **8:50am (PO20-0056)**

Three Methods to Handle Data Scarcity: Low-Resource NMT of Classical Latin Poetry

Lucas Leary (University of Puget Sound)

For this low-resource poetry generation experiment, I trained a bidirectional recurrent neural network on the poetry of Catullus and used the network to attempt to reproduce Latin gender violations. I experimented with three methods for compensating for data scarcity by manipulating the original text. In conclusion, I found that training a neural network on a dataset both forwards and reversed leads to better vocabulary use, which could be helpful for working with small data.

➢ **9:00am (PO20-0067)**

Developing a Computational Model to Diagnose Patients with Dementia

Christina Vu (Union College)

Numerous individuals in the aging population do not find out that they are susceptible to dementia until it is far too late. The time and cost involved in diagnosing a single patient may not be the most effective way to determine whether an individual is at risk. There has been numerous experimentation in using these large healthcare data depository to build predictive models that can predict and diagnose patients who are susceptible to dementia. These models utilize data that is routinely collected during primary care visits and therefore would be more efficient to implement and automate in calculating risk and diagnosing. Part of my research is to use a large hospital database which holds information about patients admitted into the critical care unit. Patients who were diagnosed with dementia were compared with patients who had dementia-like conditions and other conditions. In order to employ the databases, appropriate information has to be cleaned, extracted, and organized to ensure systematic performance for the models. For our research purposes, we used Postgresql to organize each patient's information into bigger tables holding necessary information for statistical analysis. For analysis, demographic, laboratory, and diagnostic information were analyzed to understand the nature of each patient group. 1363 patients were diagnosed with dementia (DP) and 10125 with non dementia (non-DP) out of a total of 38581 distinct adult patients admitted into critical care unit from 2001 to 2012. The most frequent diagnosis among all three groups of patients was essential hypertension.