Proceedings of the
ACM SIGKDD Workshop on
Knowledge-infused Mining and Learning for Social Impact

Editors: Manas Gaur, Alejandro Jaimes, Fatma Özcan, Srinivasan Parthasarathy, Sameena Shah, Amit Sheth & Biplav Srivastava

KiML 2020

AUGUST 24

SAN DIEGO , CA

First International Workshop on Advancing Decision making in Health, Crisis Response, and Finance

Co-located with
26th ACM Conference on
Knowledge Discovery and Data Mining
KDD 2020, San Diego, California

http://kiml2020.aiisc.ai/

Proceedings of the
ACM SIGKDD Workshop on
Knowledge-infused Mining and Learning (KiML)

These proceedings are not included in the ACM Digital Library.

KiML'20, August 24, 2020, San Diego, California, USA.

ACM SIGKDD Workshop on Knowledge-infused Mining and Learning (KiML)

Preface

Research in artificial intelligence and data science is accelerating rapidly due to an unprecedented explosion in the amount of information on the web. In parallel, we noticed immense growth in the construction and utility of the knowledge network from Google, Netflix, NSF, and NIH. However, current methods risk an unsatisfactory ceiling of applicability due to shortcomings in bringing homogeneity between knowledge graphs, data mining, and deep learning. In this changing world, retrospective studies for building state-of-the-art AI and Data science systems have raised concerns on trust, traceability, and interactivity for prospective applications in healthcare, finance, and crisis response. We believe the paradigm of knowledge-infused mining and learning would account for both pieces of knowledge that accrue from domain expertise and guidance from physical models. Further, it will allow the community to design new evaluation strategies that assess robustness and fairness across all comparable state-of-the-art algorithms.

The Workshop on Knowledge-infused Mining and Learning for Social Impact was centered around the following thematic components: (a) Data Management: includes resource management, resource discovery across heterogeneous and inconsistent data resources. (b) Data Usage: includes methods and systems for visualization, representations, reasoning, and interaction. (c) Evaluation: will bring together researchers involved at the intersection of databases, semantic web, information systems, and AI to create new approaches and tools to benefit a broad range of policymakers (e.g. mental health professions, education practitioners, emergency responders, and economists).

The workshop will bring together researchers and practitioners from both academia and industry who are interested in the creation and use of knowledge graphs in understanding online conversations on crisis response (e.g., COVID-19), public health (e.g., social network analysis for mental health insights), and finance (e.g., mining insights on the financial impact (recession, unemployment) of COVID-19 using twitter or organizational data). Additionally, we encourage researchers and practitioners from the areas of human-centered computing, interaction and reasoning, statistical relational mining and learning, intelligent agent systems, semantic social network analysis, deep graph learning, and recommendation systems.

The main program of KiML'20 consist of seven papers, selected out of thirteen submissions, covering topics related to knowledge-enabled feature elicitation, adversarial learning, crisis response, public health, and COVID-19. We sincerely thank the authors of the submissions as well as the attendees of the workshop. We wish to thank the members of our program committee for their help in selecting high-quality papers. Furthermore, we are grateful to Manuela Veloso, Sriraam Natarajan, Jose Ambite, and Pieter De Leenheer for giving keynote presentations on their recent work on Symbiotic Autonomy, Human Allied Probabilistic Learning, Biomedical Data Science, and Data Intelligence.

Manas Gaur, Alejandro Jaimes, Fatma Özcan, Srinivasan Parthasarathy,
Sameena Shah, Amit Sheth, and Biplav Srivastava
August 2020

Table of Contents

Keynote Talk 1

Symbiotic Autonomy: Knowing When and What to Learn from Experience

Manuela M. Veloso
Head, JPMorgan AI Research
Herbert A. Simon University Professor, School of Computer Science
Carnegie Mellon University
manuela.veloso@jpmchase.com

Abstract:

The talk will present work on novel human-AI interaction, in which humans and AI complement each other in their knowledge and learning. I will discuss examples in autonomous mobile service robots and in the financial domain. I will conclude with a brief discussion of multiple forms of available knowledge for AI systems that continuously learn from experience.

Bio:
Manuela M. Veloso is the Head of J.P. Morgan AI Research, which pursues fundamental research in areas of core relevance to financial services, including data mining and cryptography, machine learning, explainability, and human-AI interaction. J.P. Morgan AI Research partners with applied data analytics teams across the firm as well as with leading academic institutions globally. Professor Veloso is on leave from Carnegie Mellon University as the Herbert A. Simon University Professor in the School of Computer Science, and the past Head of the Machine Learning Department. With her students, she had led research in AI, with a focus on robotics and machine learning, having concretely researched and developed a variety of autonomous robots, including teams of soccer robots, and mobile service robots. Her robot soccer teams have been RoboCup world champions several times, and the CoBot mobile robots have autonomously navigated for more than 1,000km in university buildings. Professor Veloso is the Past President of AAAI, (the Association for the Advancement of Artificial Intelligence), and the co-founder, Trustee, and Past President of RoboCup. Professor Veloso has been recognized with multiple honors, including being a Fellow of the ACM, IEEE, AAAS, and AAAI. She is the recipient of several best paper awards, the Einstein Chair of the Chinese Academy of Science, the ACM/SIGART Autonomous Agents Research Award, an NSF Career Award, and the Allen Newell Medal for Excellence in Research. Professor Veloso earned a Bachelor and Master of Science degrees in Electrical and Computer Engineering from Instituto Superior Tecnico in Lisbon, Portugal, a Master of Arts in Computer Science from Boston University, and Master of Science and Ph.D. in Computer Science from Carnegie Mellon University. See www.cs.cmu.edu/~mmv/Veloso.html for her scientific publications.

Keynote Talk 2

Human Allied Probabilistic Learning

Sriraam Natarajan
Director, Center for Machine Learning
Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
sriraam.natarajan@utdallas.edu

Abstract:

Historically, Artificial Intelligence has taken a symbolic route for representing and reasoning about objects at a higher-level or a statistical route for learning complex models from large data. To achieve true AI, it is necessary to make these different paths meet and enable seamless human interaction. First, I briefly will introduce learning from rich, structured, complex, and noisy data. Next, I will present the recent progress that allows for more reasonable human interaction where the human input is taken as "advice" and the learning algorithm combines this advice with data. The advice can be in the form of qualitative influences, preferences over labels/actions, privileged information obtained during training, or simple precision-recall trade-off. Finally, I will outline our recent work on "closing-the-loop" where information is solicited from humans as needed that allows for seamless interactions with the human expert. While I will discuss these methods primarily in the context of probabilistic and relational learning, I will also present our results on reinforcement learning and inverse reinforcement learning.

Bio:

Dr. Sriraam Natarajan is an Associate Professor and the Director of the Center for ML at the Department of Computer Science at the University of Texas Dallas. He was previously an Associate Professor and earlier an Assistant Professor at Indiana University, Wake Forest School of Medicine, a post-doctoral research associate at the University of Wisconsin-Madison, and had graduated with his Ph.D. from Oregon State University. His research interests lie in the field of Artificial Intelligence, with emphasis on Machine Learning, Statistical Relational Learning and AI, Reinforcement Learning, Graphical Models, and Biomedical Applications. He has received the Young Investigator award from US Army Research Office, Amazon Faculty Research Award, Intel Faculty Award, XEROX Faculty Award, Verisk Faculty Award, and the IU Trustees Teaching Award from Indiana University. He is the program co-chair of SDM 2020 and ACM CoDS-COMAD 2020 conferences. He is the specialty chief editor of Frontiers in ML and AI journal, an editorial board member of MLJ, JAIR, and DAMI journals and is the electronics publishing editor of JAIR.

Keynote Talk 3

Data Intelligence in the Age of Accountability

Pieter De Leenheer
Senior Research Fellow, Harvard Business School
Co-Founder and Chief Science Officer, Collibra Inc.
pdeleenheer@hbs.edu

Abstract:

Knowledge graphs, machine learning and distributed ledgers are just a few of the emerging intelligent technologies that unlock new options to innovate business models, augment scientific knowledge and self-understanding, and enhance decision making. Data being a critical driver for intelligent systems implies machine calculation may supplant human decision making in many scenarios. The accessibility, quality and currency of data are necessary criteria to ensure these systems produce viable innovation options that can be accounted for. But are these criteria sufficient?

Bio:

Pieter is a senior research fellow at Harvard Business School and serves as adjunct faculty at Columbia University. He is a cofounder and former Chief Science Officer of Collibra, a unicorn venture in data intelligence, that spun off his PhD research on community-based ontology management. Pieter writes, teaches and advises on computing and management aspects of data innovation, accountability and citizenship. He serves as an expert to the European Commission and several governments; and as board member of several startups such as Gluetech.com and Yesse.tech. Prior to cofounding the company, Pieter was a professor at VU University of Amsterdam. He lives in New York City with his family.

Keynote Talk 4

Semantics in Biomedical Data Science

Jose Luis Ambite
Research Team Leader, Information Sciences Institute
Associate Research Professor, University of Southern California
ambite@isi.edu

Abstract:
There is an explosion of biomedical data that promises to enable novel discoveries, treatments, and the ultimate goal of personalized medicine. These data are generated in a great variety of forms, ranging from sensor data, to imaging, to genetics, and all types of clinical data. Moreover, the data are often scattered across organizations, and even for the same data type are represented in diverse structures. Thus, the need to provide a semantically consistent view, so that the data can be meaningfully analyzed is critical. I will describe core data integration and knowledge graph construction techniques, namely entity linkage and formal schema mappings, with illustrative biomedical data integration applications, highlighting some novel neural semantic similarity methods and some surprising applications of record linkage techniques, such as efficiently finding genetically related individuals. I will discuss architectures for large scale data integration and analysis, including sensor data. Finally, I will discuss how we can analyze distributed datasets when the data cannot be shared for privacy or security reasons, and thus cannot be integrated. I will describe our recent work on Heterogeneous Federated Learning that learns common neural models from siloed data.

Bio:
Dr. Jose Luis Ambite is an Associate Research Professor at the Computer Science Department, and a Research Team Leader at the Information Sciences Institute, at the University of Southern California. His core expertise is on information integration, including query rewriting under constraints, learning schema mappings, and entity linkage. Dr. Ambite research interests include databases, knowledge representation, semantic web, semantic similarity, scientific workflows, and biomedical data science. He has published widely in these topics. He regularly serves as reviewer for funding organizations, journals and major conferences. In the last years, he has focused on developing novel approaches for integration, analysis, and dissemination of biomedical and genetic data within several large NIH-funded projects, such as PRISMS-study, NIMH Repository and Genetics Resource, SchizConnect, Population Architecture using Genomics and Epidemiology, and Education Resource Discovery Index.

Textual Evidence for the Perfunctoriness of Independent Medical Reviews

Adrian Brasoveanu
abrsvn@ucsc.edu
University of California Santa Cruz
Santa Cruz, CA

Megan Moodie
mmoodie@ucsc.edu
University of California Santa Cruz
Santa Cruz, CA

Rakshit Agrawal
ragrawal@camio.com
Camio Inc.
San Mateo, CA

ABSTRACT

We examine a database of 26,361 Independent Medical Reviews (IMRs) for privately insured patients, handled by the California Department of Managed Health Care (DMHC) through a private contractor. IMR processes are meant to provide protection for patients whose doctors prescribe treatments that are denied by their health insurance (either private insurance or the insurance that is part of their worker comp; we focus on private insurance here). Laws requiring IMR were established in California and other states because patients and their doctors were concerned that health insurance plans deny coverage for medically necessary services. We analyze the text of the reviews and compare them closely with a sample of 50000 Yelp reviews [19] and the corpus of 50000 IMDB movie reviews [10]. Despite the fact that the IMDB corpus is twice as large as the IMR corpus, and the Yelp sample contains almost twice as many reviews, we can construct a very good language model for the IMR corpus using inductive sequential transfer learning, specifically ULMFiT [8], as measured by the quality of text generation, as well as low perplexity (11.86) and high categorical accuracy (0.53) on unseen test data, compared to the larger Yelp and IMDB corpora (perplexity: 40.3 and 37, respectively; accuracy: 0.29 and 0.39). We see similar trends in topic models [17] and classification models predicting binary IMR outcomes and binarized sentiment for Yelp and IMDB reviews. We also examine four other corpora (drug reviews [6], data science job postings [9], legal case summaries [5] and cooking recipes [11]) to show that the IMR results are not typical for specialized-register corpora. These results indicate that movie and restaurant reviews exhibit a much larger variety, more contentful discussion, and greater attention to detail compared to IMR reviews, which points to the possibility that a crucial consumer protection mandated by law fails a sizeable class of highly vulnerable patients.

CCS CONCEPTS

• **Computing methodologies** → **Latent Dirichlet allocation**; **Neural networks**.

KEYWORDS

AI for social good, state-managed medical review processes, language models, topic models, sentiment classification

In M. Gaur, A. Jaimes, F. Ozcan, S. Shah, A. Sheth, B. Srivastava, Proceedings of the Workshop on Knowledge-infused Mining and Learning (KDD-KiML 2020). San Diego, California, USA, August 24, 2020.
KiML'20, August 24, 2020, San Diego, California, USA,

https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Adrian Brasoveanu, Megan Moodie, and Rakshit Agrawal. 2020. Textual Evidence for the Perfunctoriness of Independent Medical Reviews. In *Proceedings of KDD Workshop on Knowledge-infused Mining and Learning (KiML'20).* , 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

1.1 Origin and structure of IMRs

Independent Medical Review (IMR) processes are meant to provide protection for patients whose doctors prescribe treatments that are denied by their health insurance – either private insurance or the insurance that is part of their workers' compensation. In this paper, we focus exclusively on privately insured patients. Laws requiring IMR processes were established in California and other states in the late 1990s because patients and their doctors were concerned that health insurance plans deny coverage for medically necessary services to maximize profit.[1]

As aptly summarized in [1], IMR is regularly used to settle disputes between patients and their health insurers over what is medically necessary or experimental/investigational care. Medical necessity disputes occur between health plans and patients because the health plan disagrees with the patient's doctor about the appropriate standard of care or course of treatment for a specific condition. Under the current system of managed care in the U.S., services rendered by a health care provider are reviewed to determine whether the services are medically necessary, a process referred to as utilization review (UR). UR is the oversight mechanism through which private insurers control costs by ensuring that only medically necessary care, covered under the contractual terms of a patient's insurance plan, is provided. Services that are not deemed medically necessary or fall outside a particular plan are not covered.

Procedures or treatment protocols are deemed experimental or investigational because the health plan – but not necessarily the patient's doctor, who in many cases has enough clinical confidence in a treatment to order it – considers them non-routine medical care, or takes them to be scientifically unproven to treat the specific condition, illness, or diagnosis for which their use is proposed.

It is important to realize that the IMR process is usually the third and final stage in the medical review process. The typical progression is as follows. After in-person and possibly repeated examination of the patient, the doctor recommends a treatment,

[1] For California, see the Friedman-Knowles Act of 1996, requiring California health plans to provide external independent medical review (IMR) for coverage denials. As of late 2002, 41 states and the District of Columbia had passed legislation creating an IMR process. In 34 of these states, including California, the decision resulting from the IMR is binding to the health plan. See [1, 15] for summaries of the political and legal history of the IMR system, and [2] for an early partial survey of the DMHC IMR data.

which is then submitted for approval to the patient's health plan. If the treatment is denied in this first stage, both the doctor and the patient may file an appeal with the health plan, which triggers a second stage of reviews by the health-insurance provider, for which a patient can supply additional information and a doctor may engage in what is known as a "peer to peer" discussion with a health-insurance representative. If these second reviews uphold the initial denial, the only recourse the patient has is the state-regulated IMR process, and per California law, an IMR grievance form (and some additional information) is included with the denial letter.

An IMR review must be initiated by the patient and submitted to the California Department of Managed Health Care (DMHC), which manages IMRs for privately-insured patients. Motivated treating physicians may provide statements of support for inclusion in the documentation provided to DMHC by the patient, but in theory the IMR creates a new relationship of care between the reviewing physician(s) hired by a private contractor on behalf of DMHC, and the patient in question. The reviewing physicians' decision is supposed to be made based on what is in the best interest of the patient, not on cost concerns. It is this relation of care that constitutes the consumer protection for which IMR processes were legislated. Understandably, given that the patients in question may be ill or disabled or simply discouraged by several layers of cumbersome bureaucratic processes, there is a very high attrition from the initial review to the final, IMR, stage. That is, only the few highly motivated and knowledgeable patients – or the extremely desperate – get as far as the IMR process.

The IMR process is regulated by the state, but it is actually conducted by a third party. At this time (2019), the provider in California and several other states across the US is MAXIMUS Federal Services, Inc.[2] The costs associated with the IMR review, at least in California, are covered by health insurers. It is DMHC's and MAXIMUS's responsibility to collect all the documentation from the patient, the patient's doctor(s) and the health insurer. There are no independent checks that all the documentation has actually been collected, however, and patients do not see a final list of what has been provided to the reviewer prior to the IMR decision itself (a *post facto* list of file contents is mailed to patients along with the final, binding, decision; it is unclear what recourse a patient may have if they find pertinent information was missing from the review file). Once the documentation is assembled, MAXIMUS forwards it to anywhere from one to three reviewers, who remain anonymous, but are certified by MAXIMUS to be appropriately credentialed and knowledgeable about the treatment(s) and condition(s) under review. The reviewer submits a summary of the case, and also a rationale and evidence in support of their decision, which is a binary *Upheld/Overturned* decision about the medical service. IMR reviewers do not enter a consultative relationship with the patient, doctor or health plan – they must render an uphold/overturn decision based solely on the provided medical records. However, as noted above, they are in an implied relationship of care to the patient, a point to which we return in the Discussion section below (§4).

While insurance carriers do not provide statistics about the percentage of requested treatments that are denied in the initial stage, looking at the process as a whole, a pattern of service denial aimed to maximize profit, rather than simply maintain cost effectiveness, seems to emerge. Typically, the argument for denial contends that the evidence for the beneficial effects of the treatment fails the prevailing standard of scientific evidence. This prevailing standard invoked by IMR reviewers is usually randomized control trials (RCTs), which are expensive, time-consuming trials that are run by large pharmaceutical companies only if the treatment is ultimately estimated to be profitable.

RCTs, however, have known limits: they "require minimal assumptions and can operate with little prior knowledge [which] is an advantage when persuading distrustful audiences, but it is a disadvantage for cumulative scientific progress, where prior knowledge should be built upon, not discarded." [3] Inflexibly applying the RCT "gold standard" in the IMR process is often a way to ignore the doctors' knowledge and experience in a way that seems superficially well-reasoned and scientific. "RCTs can play a role in building scientific knowledge and useful predictions" – and we add, treatment recommendations – "only [...] as part of a cumulative program, [in combination] with other methods." [3]

Notably, the experimental/investigational category of treatments that get denied often includes promising treatments that have not been fully tested in clinical RCTs – because the treatment is new or the condition is rare in the population, so treatment development costs might not ultimately be recovered. Another common category of experimental/investigational denials involves "off-label" drug uses, that is, uses of FDA-approved pharmaceuticals for a purpose other than the narrow one for which the drug was approved.

1.2 Main argument and predictions

Recall that these *'experimental' treatments or off-label uses are recommended by the patient's doctor*, and therefore their potential benefits are taken to outweigh their possible negative effects. The recommending doctor is likely very familiar with the often lengthy, tortuous and *highly specific medical history of the patient*, and with the list of 'less experimental' treatments that have been proven unsuccessful or have been removed from consideration for patient-specific reasons. It is also important to remember that many *rare conditions have no "on-label" treatment options available*, since expensive RCTs and treatment approval processes are not undertaken if companies do not expect to recover their costs, which is likely if the potential 'market' is small (few people have the rare condition).

Therefore, our main line of argumentation is as follows.

- Since IMRs are the final stage in a long bureaucratic process in which health insurance companies keep denying coverage for a treatment repeatedly recommended by a doctor as medically necessary, we expect that the issue of medical necessity is non-trivial when that specific patient and that specific treatment are carefully considered.
- We should therefore expect the text of the IMRs, which justifies the final determination, to be highly individualized and argue for that final decision (whether congruent with the health plan's decision or not) in a way that involves the particulars of the treatment and the particulars of the patient's medical history and conditions.

Thus, we expect a reasoned, thoughtful IMR to *not be highly generic and templatic / predictable* in nature. For instance, legal

[2] https://www.maximus.com/capability/appeals-imr

documents may be highly templatic as they discuss the application of the same law or policy across many different cases, but a response carefully considering the specifics of a medical case reaching the IMR stage is not likely to be similar to many other cases. We only expect high similarity and 'templaticity' for IMR reviews if they are reduced to a more or less automatic application of some prespecified set of rules (rubber-stamping).

1.3 Main results, and their limits

Concomitantly with this quantitative study, we conducted preliminary qualitative research with a focus on pain management and chronic conditions. We investigated the history of the IMR process, in addition to having direct experience with it. We had detailed conversations with doctors in Northern California and on private social media groups formed around chronic conditions and pain management. This preliminary research reliably points towards the possibility that IMR reviews are perfunctory, and that this crucial consumer protection mandated by law seems to fail for a sizeable class of highly vulnerable patients. In this paper, we focus on the text of the IMR decisions and attempt to quantify the evidence for the perfunctoriness of the IMR process that they provide.

The text of the IMR findings does not provide unambiguous evidence about the quality and appropriateness of the IMR process. If we had access to the full, anonymized patient files submitted to the IMR reviewers (in addition to the final IMR decision and the associated text), we might have been able to provide much stronger evidence that IMRs should have a significantly higher percentage of overturns, and that the IMR process should be improved in various ways, e.g., (*i*) patients should be able to check that all the relevant documentation has been collected and will be reviewed, and (*ii*) the anonymous reviewers should be held to higher standards of doctor-patient care. At the very least, one would want to compare the reports/letters produced by the patient's doctor(s) and the IMR texts. However, such information is not available and there are no visible signs suggesting potential availability in the near future. The information that is made available by DMHC constitutes the IMR decision – whether to uphold or overturn the health plan decision –, the anonymized decision letter, and information about the requested treatment category (also available in the letter). We, therefore, had to limit ourselves to the text of the DMHC-provided IMR findings in our empirical analysis.

A qualitative inspection of the corpus of IMR decisions made available by the California DMHC site as of June 2019 (a total of 26,631 cases spanning the years 2001-2019) indicates that the reviews – as documented in the text of the findings – focus more on the review procedure and associated legalese than on the actual medical history of the patient and the details of the case. For example, decisions for chronic pain management seem to mostly rubber-stamp the Medical Treatment Utilization Schedule (MTUS) guidelines, with very little consideration of the rarity of the underlying condition(s) (see our comments about RCTs above), or a thoughtful evaluation of the risk/benefit profile of the denied treatment relative to the specific medical history of the patient (assuming this history was adequately documented to begin with).

The goal in this paper is to investigate to what extent Natural Language Processing (NLP) / Machine Learning (ML) methods that are able to extract insights from large corpora point in the same direction, thus mitigating cherry-picking biases that are sometimes associated with qualitative investigations. In addition to the IMR text, we perform a comparative study with additional English-language datasets in an attempt to eliminate data-specific and problem-specific biases.

- We analyze the text of the IMR reviews and compare them with a sample of 50,000 Yelp reviews [19] and the corpus of 50,000 IMDB movie reviews [10].
- As the size of data has significant consequences for language-model training, and NLP/ML models more generally, we expect models trained on the Yelp and IMDB corpora to outperform models trained on the IMR corpus, given that the IMDB corpus is twice as large as the IMR corpus, and the Yelp samples contain almost twice as many reviews.
- In this paper, we instead demonstrate that we were able to construct a very good language model for the IMR corpus using inductive sequential transfer learning, specifically ULMFiT [8], as measured by the quality of text generation.
- In addition, the model achieves a much lower perplexity (11.86) and a higher categorical accuracy (0.53) on unseen test data, compared to models trained on the larger Yelp and IMDB corpora (perplexity: 40.3 and 37, respectively; categorical accuracy: 0.29 and 0.39).
- We see similar trends in topic models [17] and classification models predicting binary IMR outcomes and binarized sentiment for Yelp and IMDB reviews.

These results indicate that movie and restaurant reviews exhibit a much larger variety, more contentful discussion, and greater attention to detail compared to IMR reviews. In an attempt to mitigate confirmation bias, as well as potentially significant register differences between IMRs and movie or restaurant reviews, we examine four additional corpora: drug reviews [6], data science job postings [9], legal case summaries [5] and cooking recipes [11]. These specialized-register corpora are potentially more similar to IMRs than IMDB or Yelp: the texts are more likely to be highly similar, include boilerplate text and have a templatic/standardized structure. We find that predictability of IMR texts, as measured by language-model perplexity and categorical accuracy, is higher than all the comparison datasets by a good margin.

Based on these empirical comparisons, we conclude that we have strong evidence that the IMR reviews are perfunctory and, therefore, that a crucial consumer protection mandated by law seems to fail for a sizeable class of highly vulnerable patients. The paper is structured as follows. In Section 2, we discuss the datasets in detail, with a focus on the nature and characteristics of the IMR data. In Section 3, we discuss the models we use to analyze the IMR, Yelp and IMDB datasets, as well as the four auxiliary corpora (drug reviews, data science jobs, legals cases and recipes). The section also compares and discusses the results of these models. Section 4 puts all the results together into an argument for the perfunctoriness of the IMRs. Section 5 concludes the paper and outlines directions for future work.

2 THE DATASETS

2.1 The IMR dataset

The IMR dataset was obtained from the DMHC website in June 2019[3] and was minimally preprocessed. It contains 26,361 cases / observations and 14 variables, 4 of which are the most relevant:

- TreatmentCategory: the main treatment category;
- ReportYear: year the case was reported;
- Determination: indicates if the determination was upheld or overturned;
- Findings: a summary of the case findings.

The top 14 treatment categories (with percentages of total $\geq 2\%$), together with their raw counts and percentages are provided in Table 1.

Table 1: Top 14 treatment categories

TreatmentCategory	Case count	% of total
Pharmacy	6480	25%
Diag Imag & Screen	4187	16%
Mental Health	2599	10%
DME	1714	7%
Gen Surg Proc	1227	5%
Orthopedic Proc	1173	5%
Rehab/ Svc - Outpt	1157	4%
Cancer Care	1029	4%
Elect/Therm/Radfreq	828	3%
Reconstr/Plast Proc	825	3%
Autism Related Tx	767	3%
Emergency/Urg Care	582	2%
Diag/ MD Eval	573	2%
Pain Management	527	2%

The breakdown of cases by patient gender (not recorded for all cases) is as follows: Female – 14823 (56%), Male – 10836 (41%), Other – 11 (0.0004%).

The breakdown by determination (the outcome of the IMR) is: Upheld – 14309 (54%), Overturned – 12052 (46%).

The outcome counts and percentages by year are provided in Table 2. The number of cases for 2019 include only the first 5 months of the year plus a subset of June 2019.

Interestingly, the DMHC website featured a graphic in June 2019 (Figure 1) that reports the percentage of Overturned outcomes to be 64%, a figure that does not accord with any of our data summaries. We intend to follow up on this issue and see if the DMHC can share their data-analysis pipeline so that we can pinpoint the source(s) of this difference.

Given that our main goal here is to investigate the text of the IMR findings and its predictiveness with respect to IMR outcomes, we provide some general properties of this corpus. The histogram of word counts for the IMR findings (the text associated with each case) is provided in Figure 2. There are 26,361 texts, with a total of 5,584,280 words. Words are identified by splitting texts on white space (sufficient for our purposes here). The mean length of a text is 211.84 words, with a standard deviation (SD) of 120.58.

[3]https://data.chhs.ca.gov/dataset/independent-medical-review-imr-determinations-trend.

Table 2: Outcome counts and percentages by year

ReportYear	Total # of cases	Overturned	Upheld
2001	28	7 (25%)	21
2002	695	243 (35%)	452
2003	738	280 (38%)	458
2004	788	305 (39%)	483
2005	959	313 (33%)	646
2006	1080	442 (41%)	638
2007	1342	571 (43%)	771
2008	1521	678 (45%)	843
2009	1432	641 (45%)	791
2010	1453	661 (45%)	792
2011	1435	684 (48%)	751
2012	1203	589 (49%)	614
2013	1197	487 (41%)	710
2014	1433	549 (38%)	884
2015	2079	1070 (51%)	1009
2016	3055	1714 (56%)	1341
2017	2953	1391 (47%)	1562
2018	2545	1218 (48%)	1327
2019	425	209 (49%)	216

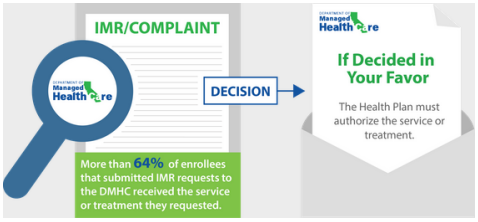


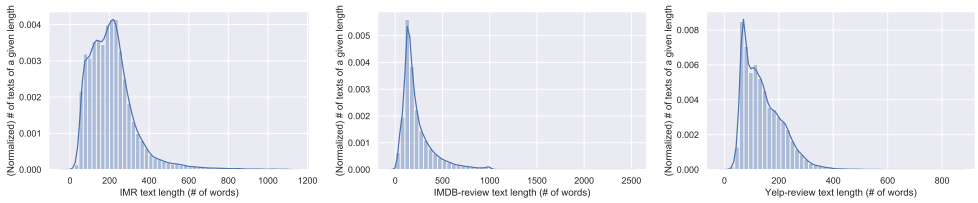
Figure 1: % Overturned claimed on DMHC site (June 2019)

2.2 The comparison datasets

As comparison datasets, we use the IMDB movie-review dataset [10], which has 50,000 reviews and a binary positive/negative sentiment classification associated with each review. This dataset will be particularly useful as a baseline for our ULMFiT transfer-learning language models (and subsequent transfer-learning classification models), where we show that we obtain results for the IMDB dataset that are similar to the ones in the original ULMFiT paper [8].

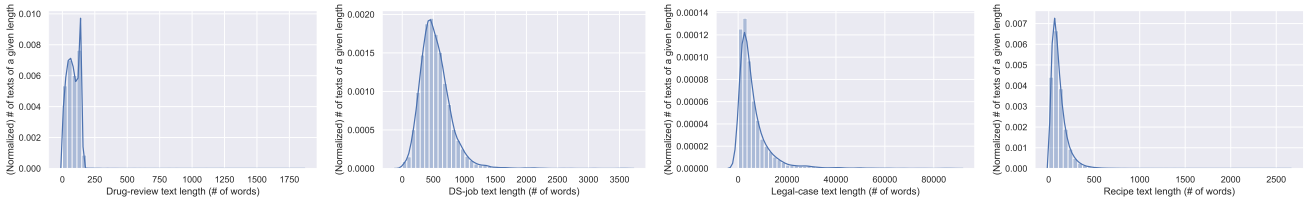
There are 50,000 movie reviews in the IMDB dataset, evenly split into negative and positive reviews. The histogram of text lengths for IMDB reviews is provided in Figure 2. The reviews contain a total of 11,557,297 words. The mean length of a review is 231.15 words, with an SD of 171.32.

We select a sample of 50,000 Yelp (mainly restaurant) reviews [19], with associated binarized negative/positive evaluations, to provide a comparison corpus intermediate between our DMHC dataset and the IMDB dataset. From a total of 560,000 reviews (evenly split between negative and positive), we draw a weighted random sample with the weights provided by the histogram of text lengths for the IMR corpus. The resulting sample contains 25,809 (52%) negative reviews and 24,191 (48%) positive reviews. The histogram of text lengths for Yelp reviews is also provided in Figure 2. The reviews contain a total of 7,038,467 words. The mean length of a review is 140.77 words, with an SD of 71.09.



(a) IMR **(b) IMDB** **(c) Yelp**

Figure 2: Histograms of text lengths (numbers of words per text) for the IMR, IMDB and Yelp corpora



(a) Drug Reviews **(b) DS Jobs** **(c) Legal cases** **(d) Recipes**

Figure 3: Histograms of text lengths (numbers of words per text) for the auxiliary datasets

2.3 Four auxiliary datasets

We will also analyze four other specialized-register corpora: drug reviews [6], data science (DS) job postings [9], legal case reports [5] and cooking recipes [11]. The modeling results for these specialized-register corpora will enable us to better contextualize and evaluate the modeling results for the IMR, IMDB and Yelp corpora, since these four auxiliary datasets might be seen as more similar to the IMR corpus than movie or restaurant reviews. The drug-review corpus contains reviews of pharmaceutical products, which are closer in subject matter to IMRs than movie/restaurant reviews. The other three corpora are all highly specialized in register, just like the IMRs, with two of them (DS jobs and legal cases) particularly similar to the IMRs in that they involve templatic texts containing information aimed at a specific professional sub-community.

These four corpora are very different from each other and from the IMR corpus in terms of (*i*) the number of texts that they contain and (*ii*) the average text length (number of words per text). Because of this, there was no obvious way to sample from them and from the IMR, IMDB and Yelp corpora in such a way that the resulting samples were both roughly comparable with respect to the total number of texts and average text length, and also large enough to obtain reliable model estimates. We therefore analyzed these four corpora as a whole.

The drug-review corpus includes 132,300 drugs reviews – more than the double the number of texts in the IMDB and Yelp datasets, and more than 4 times the number of texts in the IMR dataset. From the original corpus of 215,063 reviews, we only retained the reviews associated with a rating of 10, which we label as *positive* reviews, and a rating of 1 through 5, which we label as *negative* reviews.[4]

The histogram of text lengths for drug reviews is provided in Figure 3. The reviews contain a total of 11,015,248 words, with a mean length of 83.26 words per review (significantly shorter than the IMR/IMDB/Yelp texts) and an SD of 45.73.

The DS corpus includes 6,953 job postings (about a quarter of the texts in the IMR corpus), with a total of 3,731,051 words. The histogram of text lengths is provided in Figure 3. The mean length of a job posting is 536.61 words (more than twice as long as the IMR/IMDB/Yelp texts), with an SD of 254.06.

There are 3,890 legal-case reports (even fewer than DS job postings), with a total of 25,954,650 words (about 5 times larger than the IMR corpus). The histogram of text lengths for the legal-case reports is provided in Figure 3. The mean length of a report is 6,672.15 words (a degree of magnitude longer than IMR/IMDB/Yelp), with a very high SD of 11,997.98.

Finally, the recipe corpus includes more than 1 million texts: there are 1,029,719 recipes, with a total of 117,563,275 words (very large compared to our other corpora). The histogram of text lengths for the recipes is provided in Figure 3. The mean length of a recipe is 114.17 words (close to the length of a drug review, and roughly half of an IMR), with an SD of 90.54.

3 THE MODELS

In this section, we analyze the text of the IMR findings and its predictiveness with respect to IMR outcomes. We systematically compare these results with the corresponding ones for the IMDB and Yelp corpora. The datasets were split into training (80%), validation (10%) and test (10%) sets. Test sets were only used for the final model evaluation.

[4]We did this so that we have a fairly balanced dataset (68,005 *positive* drug reviews and 64,295 *negative* reviews) to estimate classification models like the ones we report for the IMR, IMDB and Yelp corpora in the next section. For completeness, the drug-review classification results on previously unseen test data are as follows: logistic regression accuracy: 77.89%; accuracy of multilayer perceptron with a 1,000-unit hidden layer and a ReLU non-linearity: 83.18%; ULMFiT classification model accuracy: 96.12%.

We start with baseline classification models (logistic regressions and logistic multilayer perceptrons with one hidden layer) to establish that the reviews in all three datasets under consideration are highly predictive of the associated binary outcomes. Once the predictiveness, hence, relevance, of the text is established, we turn to an in-depth analysis of the texts themselves by means of topic and language models. We see that the text of the IMR reviews is significantly different (more predictable, less diverse / contentful) when compared to movie and restaurant reviews. We then turn to a final set of classification models that leverage transfer learning from the language models to see how predictive the texts can really be with respect to the associated binary outcomes. Finally, we report the results of estimating language models for the 4 auxiliary datasets introduced in the previous section.

The main conclusion of this extensive series of models is that the IMR corpus is an outlier, and it would be easy to make the IMR process fully automatic: it is pretty straightforward to train models that generate high-quality, realistic IMR reviews and generate binary decisions that are very reliably associated with these reviews. In contrast, movie and restaurant reviews produced by unpaid volunteers (as well as the 4 auxiliary datasets) exhibit more human-like depth, sophistication and attention to detail, so current NLP models do not perform as well on them.

3.1 Classification models

We regress outcomes (Upheld/Overturned for IMR or negative/positive sentiment for IMDB/Yelp) against the text of the corresponding findings / reviews. For the purposes of these basic classification models, as well as the topics models discussed in the following subsection, the texts were preprocessed as follows. First, we removed stop words; for the IMR dataset, we also removed the following high-frequency words: *patient*, *treatment*, *reviewer*, *request*, *medical* and *medically*, and for the IMDB dataset, we also removed the words *film* and *movie*. After part-of-speech tagging, we retained only nouns, adjectives, verbs and adverbs, since lexical meanings provide the most useful information for logistic (more generally, feed-forward) models and topic models. The resulting dictionary for the IMR dataset had 23,188 unique words. We ensured that the dictionaries for the IMDB and Yelp datasets were also between 23,000 and 24,000 words by eliminating infrequent words. Bounding the dictionaries for each dataset to a similar range helps mitigate dataset-specific modeling biases: having differently-sized vocabularies leads to differently-sized parameter spaces for the models.

We extracted features by converting each text into sparse bag-of-words vectors of dictionary length, which recorded how many times each token occurred in the text. These feature representations were the input to all the classifier models we consider in this subsection. The multilayer perceptron model had a single hidden layer with 1,000 units and a ReLU non-linearity. The classification accuracies on the test data for all three datasets are provided in Table 3.

Table 3: Classification accuracy for basic models

	IMR	IMDB	Yelp
LOGISTIC REGRESSION	90.75%	86.30%	87.62%
MULTILAYER PERCEPTRON	90.94%	87.14%	88.92%

We see that the text of the findings / reviews is highly predictive of the associated binary outcomes, with the highest accuracy for the IMR dataset despite the fact that it contains half the observations of the other two data sets. We can therefore turn to a more in-depth analysis of the texts to understand what kind of textual justification is used to motivate the IMR binary decisions. To that end, we examine and compare the results of two unsupervised/self-supervised types of models: topic models and language models.

3.2 Topic models

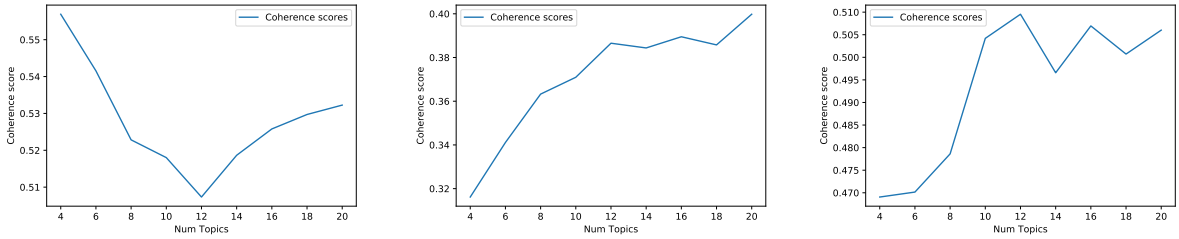
Topic modeling [17] is an unsupervised method that distills semantic properties of words and documents in a corpus in terms of probabilistic topics. The most widespread measure for topic model evaluation is the coherence score [14]. Typically, as we increase the number of topics from very few, say, 4 topics, to more of them, we see an increase in coherence score that tends to level out after a certain number of topics. When modeling the IMDB and Yelp datasets, we see exactly this behavior, as shown in Figure 4.

In contrast, the 4-topic model has the highest coherence score (0.56) for the IMR data set, also shown in Figure 4. Furthermore, as we add more topics, the coherence score drops. As the word clouds for the 4-topic model in Figure 5 show, these 4 topics mostly reflect the legalese associated with the IMR review procedure and very little, if anything, of the treatments and conditions that were the main point of the review. In contrast, the corresponding high-scoring topic models for the IMDB and Yelp datasets reflect actual features of movies, e.g., family-life movies, westerns, musicals etc., or breakfast/lunch places, restaurants, shops, bars, hotels etc.

Recall that IMRs are the legally-mandated last resort for patients seeking treatments (usually) ordered by their doctors, and which their health plan refuses to cover. The reviews are conducted exclusively based on documentation. Putting aside the fact that it is unclear how much effort is taken to ensure that the documentation is complete, especially for patients with extensive and complicated health records, we see that relatively little specific information about a patients' medical history, condition(s), or the recommended treatments are reflected in the text of these decisions. The text seems to consist largely of legalese about the IMR process, the health plan / providers, basic demographic information about the patient, and generalities about the medical service or therapy requested for the enrollee's condition.

3.3 Language models with transfer learning

Language models, specifically using neural networks, are usually recurrent-network or transformer based architectures designed to learn textual distributional patterns in an unsupervised or self-supervised manner. Recurrent-network models – on which we focus here – commonly use Long Short-Term Memory (LSTM) [7] "cells," which are able to learn long-term dependencies in sequences. Representing text as a sequence of words, language models build rich representations of the words, sentences, and their relations within a certain language. We estimate a language model for the IMR corpus using inductive sequential transfer learning, specifically ULMFiT [8]. Just as [8], we use the AWD-LSTM model [12], a vanilla LSTM with 4 kinds of dropout regularization, embedding size of 400, 3 LSTM layers (1,150 units per layer), and a BPTT of size 70.



(a) IMR (b) IMDB (c) Yelp

Figure 4: Coherence scores for topic models (x-axis: number of topics; y-axis: coherence score)

Figure 5: Word clouds for the 4-topic IMR model

The AWD-LSTM model is pretrained on Wikitext-103 [13], consisting of 28, 595 preprocessed Wikipedia articles, with a total of 103 million words. This pretrained model is fairly simple (no attention, skip connections etc.), and the pretraining corpus is of modest size.

To obtain our final language models for the IMR, IMDB and Yelp corpora, we fine-tune the pretrained AWD-LSTM model using discriminative [18] and slanted triangular [8, 16] learning rates. We do the same kind of minimal text preprocessing as in [8].

The perplexity and categorical accuracy for the 3 language models are provided in Table 4. The perplexity for the IMR findings is much lower than for the IMDB / Yelp reviews, and the language model can correctly guess the next word more than half the time.

Table 4: Language-model perplexity and categ. accuracy

	IMR	IMDB	Yelp
PERPLEXITY	11.86	36.96	40.3
CATEGORICAL ACCURACY	53%	39%	29%

The IMR language model can generate high quality and largely coherent text, unlike the IMDB / Yelp models. Two samples of generated text are provided below (the 'seed' text is boldfaced).

- **The issue in this case is** whether the requested partial hospitalization program (PHP) services are medically necessary for treatment of the patient 's behavioral health condition . The American Psychiatric Association (APA) treatment guidelines for patients with eating disorders also consider PHP acute care to be the most appropriate setting for treatment , and suggest that patients should be treated in the least restrictive setting which is likely to be safe and effective . The PHP was initially recommended for patients who were based on their own medical needs , but who were
- **The patient was admitted** to a skilled nursing facility (SNF) on 12 / 10 / 04 . The submitted documentation states the patient was discharged from the hospital on 12 / 22 / 04 . The following day the patient 's vital signs were stable . The patient had been ambulating to the community with assistance with transfers , but has not had any recent medical or rehabilitation therapy . The patient had no new medical problems and was discharged in stable condition . The patient has requested reimbursement for the inpatient acute rehabilitation services provided

We see that the IMR language model is highly performant, despite the simple model architecture we used, the modest size of the pretraining corpus, and the small size of the IMR corpus. The quality of the generated text is also very high, particularly given all these limitations.

3.4 Classification with transfer learning

We further fine-tune the language models discussed in the previous subsection to train classifiers for the three datasets. Following [4, 8], we gradually unfreeze the classifier models to avoid catastrophic forgetting.

The results of evaluating the classifiers on the withheld test sets are provided in Table 5. Despite the fact that the IMR dataset contains half of the classification observations of the other two datasets, we obtain the highest level of accuracy when predicting binary Upheld/Overturned decisions based on the text of the IMR findings.

Table 5: Accuracy for transfer-learning classifiers

	IMR	IMDB	Yelp
CLASSIFICATION ACCURACY	97.12%	94.18%	96.16%

Table 6: Comparison of language models across all datasets. Best performing metrics are boldfaced.

Dataset	Perplexity	Categorical Accuracy
IMR reviews	**11.86**	**0.53**
Legal cases	18.17	0.43
DS Jobs	22.14	0.41
Drug reviews	25.06	0.36
Recipes	29.56	0.39
IMDB	36.96	0.39
Yelp	40.3	0.29

3.5 Models for auxiliary corpora

We also estimated topic and language models for the 4 auxiliary corpora (drug reviews, DS jobs, legal cases and cooking recipes). The associations between coherence scores and number of topics for these 4 corpora was similar to the ones plotted in Figure 4 above for the IMDB and Yelp corpora. For all 4 auxiliary corpora, the best topic models had at least 14 topics, often more, with coherence scores above 0.5. The quality of the topics was also high, with intuitively coherent and contentful topics (just like IMDB / Yelp).

The perplexity and accuracy of the ULMFiT language models on previously-withheld test data are provided in Table 6, which contains the results for all the 7 datasets under consideration in this paper. We see that the predictability of the IMR corpus, as reflected in its perplexity and categorical accuracy scores, is still clearly higher than the 4 auxiliary corpora. The perplexity of the legal-case corpus (18.17) is somewhat close to the IMR perplexity (11.86), but we should remember that the legal-case corpus is about 5 times larger than the IMR corpus. Furthermore, the legal-case categorical accuracy of 43% is still substantially lower than the IMR accuracy of 53%. Notably, even the recipe corpus, which is about 20 times larger than the IMR corpus ($\approx$ 117.5 vs. $\approx$ 5.5 million words) does not have test-set scores similar to the IMR scores.

The results for these 4 auxiliary corpora indicate that the IMR corpus is an outlier, with very highly templatic and generic texts.

4 DISCUSSION

The models discussed in the previous section show that language-model learning is significantly easier for IMRs compared to the other 6 corpora. As can be seen in Table 6, perplexity in the language model for IMR reviews is clearly lower than even legal cases, for which we expect highly templatic language and high similarity between texts. This pattern can be clearly observed in Figure 6, with the IMR corpus clearly at the very end of the high-to-low predictability spectrum.

One would not expect such highly predictable texts in an ideal scenario, where each medical review is thorough, and each decision is accompanied by strong medical reasoning relying on the specifics of the case at hand, and based on an objective physician's, or team of physicians', opinion as to what is in the patient's best interest. Arguably, these medically complex cases are as diverse as Hollywood blockbusters or fashionable restaurants – the patients themselves certainly experience them as unique and meaningful –, and their reviews should be similarly diverse, or at most as templatic as a job posting or a cooking recipe. We wouldn't expect

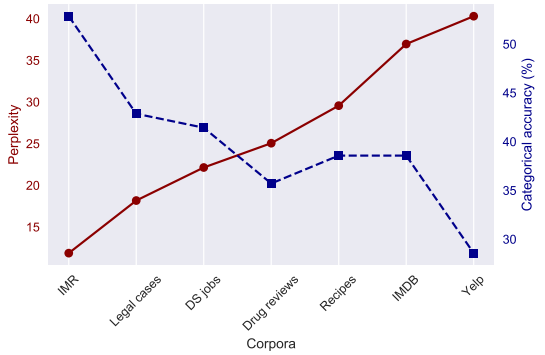


Figure 6: Comparison of language-model perplexity and categorical accuracy across all the datasets.

these medical reviews to be so much more predictable and generic than less socially consequential reviews of movies and restaurants.

What are the ethical and potentially legal consequences of these findings? First, while state legislators assume we have strong health-insurance related consumer protections in place, an image DMHC goes to great lengths to promote, we find the reviews to be upholding insurance plan denials at rates that exceed what one might expect, given that the treatments in question are frequently being ordered by a treating physician, and that the IMR process is the last stage in a bureaucratically laborious (hence high-attrition) process of appealing health-plan denials.

Second, given that the IMR process creates an implied relation of care between the reviewers hired by MAXIMUS and the patient – since reviewers are, after all, being entrusted with the best interests of the patient without regard to cost –, one can hardly say that they are fulfilling their obligations as doctors to their patient with such seemingly rote, perfunctory reviews.

Third, if IMR processes were designed to make sure that (*i*) treatment decisions are being made by doctors, not by profit-driven businesses, and (*ii*) insurance companies cannot welch on their responsibilities to plan members, one must wonder whether prescribing physicians are wrong more than half the time. Do American doctors really order so many erroneous, medically unnecessary treatments and medications? If so, how is it possible that they are so committed and confident in them that they are willing to escalate the appeal process all the way to the state-managed IMR stage? Or is it that IMRs often serve as a final rubber stamp for health-insurance plan denials, failing their stated mission of protecting a vulnerable population?

We end this discussion section by briefly reflecting on the way we used ML/NLP methods for social good problems in this paper. Overwhelmingly, the social-good applications of these methods and models seem to be predictive in nature: their goal is to improve the outcomes of a decision-making process, and the improvement is evaluated according to various performance-related metrics. An important class of metrics that are currently being developed have to do with ethical, or 'safe,' uses of ML/AI models.

In contrast, our use of ML models in this paper was analytical, with the goal of extracting insights from large datasets that enable us to empirically evaluate how well an established decision-making

process with high social impact functions. Data analysis of this kind, more akin to hypothesis testing than to predictive modeling, is in fact one of the original uses of statistical models / methods.

Unfortunately, using ML models in this way does not straightforwardly lead to plots showing how ML models obviously improve metrics like the efficiency or cost of a process. We think, however, that there are as many socially beneficial opportunities for this kind of data-analysis use of ML modeling as there are for its predictive uses. The main difference between them seems to be that the data-analysis uses do not lead to more-or-less immediately measurable products. Instead, they are meant to become part of a larger argument and evaluation of a socially and politically relevant issue, e.g., the ethical status of current health-insurance related practices and consumer protections discussed here. What counts as 'success' when ML models are deployed in this way is less immediate, but could provide at least as much social good in the long run.

5 CONCLUSION AND FUTURE WORK

We examined a database of 26,361 IMRs handled by the California DMHC through a private contractor. IMR processes are meant to provide protection for patients whose doctors prescribe treatments that are denied by their health insurance.

We found that, in a majority of cases, IMRs uphold the health insurance denial, despite DMHC's claim to the contrary. In addition, we analyzed the text of the reviews and compared them with a sample of 50,000 Yelp reviews and the IMDB movie review corpus. Despite the fact that these corpora are basically twice as large, we can construct a very good language model for the IMR corpus, as measured by the quality of text generation, as well as its low perplexity and high categorical accuracy on unseen test data. These results indicate that movie and restaurant reviews exhibit a much larger variety, more contentful discussion, and greater attention to detail compared to IMR reviews, which seem highly templatic and perfunctory in comparison. We see similar trends in topic models and classification models predicting binary IMR outcomes and binarized sentiment for Yelp and IMDB reviews.

These results were further confirmed by topic and language models for four other specialized-register corpora (drug reviews, data science job postings, legal-case reports and cooking recipes).

We are in the process of extending our datasets with (*i*) workers' comp cases from California and (*ii*) private insurance cases from other states. This will enable us to investigate if the reviews for workers' comp cases are substantially different from the DMHC IMR data (the percentage of upheld decisions is much higher for workers' comp: $\approx$ 90%), as well as if the reviews vary substantially across states.

Another direction for future work is to follow up on our preliminary qualitative research with a survey of patients that have experienced the IMR process to see if these patients agree with the DMHC-promoted message that the IMR process provides strong consumer protection against unjustified health-plan denials. This could also enable us to verify if the medical documentation collected during the IMR process is complete and actually taken into account when the decision is made.

The ultimate upshot of this project would be a list of recommendations for the improvement of the IMR process, including but not limited to (*i*) adding ways for patients to check that all the relevant documentation has been collected and will be reviewed, and (*ii*) identifying ways to hold the anonymous reviewers to higher standards of doctor-patient care.

ACKNOWLEDGMENTS

We are grateful to four KDD-KiML anonymous reviewers for their comments on an earlier version of this paper. We gratefully acknowledge the support of the NVIDIA Corporation with the donation of two Titan V GPUs used for this research, as well as the UCSC Office of Research and The Humanities Institute for a matching grant to purchase additional hardware. The usual disclaimers apply.

REFERENCES

[1] Leatrice Berman-Sandler. 2004. Independent Medical Review: Expanding Legal Remedies to Achieve Managed Care Accountability. *Annals Health Law* 13 (2004).
[2] Kenneth H. Chuang, Wade M. Aubry, and R. Adams Dudley. 2004. Independent Medical Review Of Health Plan Coverage Denials: Early Trends. *Health Affairs* 23, 6 (2004), 163–169. https://doi.org/10.1377/hlthaff.23.6.163
[3] Angus Deaton and Nancy Cartwright. 2018. Understanding and misunderstanding randomized controlled trials. *Social Science and Medicine* 210 (2018), 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005
[4] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 1615–1625. https://doi.org/10.18653/v1/D17-1169
[5] Filippo Galgani and Achim Hoffmann. 2011. LEXA: Towards Automatic Legal Citation Classification. In *AI 2010: Advances in Artificial Intelligence*, Jiuyong Li (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 445–454.
[6] Felix Gräundefineder, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning *(DH '18)*. Association for Computing Machinery, New York, NY, USA, 121–125. https://doi.org/10.1145/3194658.3194677
[7] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
[8] Jeremy Howard and Sebastian Ruder. 2018. Fine-tuned Language Models for Text Classification. *CoRR* abs/1801.06146 (2018). arXiv:1801.06146 http://arxiv.org/abs/1801.06146
[9] Shanshan Lu. 2018. Data Scientist Job Market in the U.S. https://www.kaggle.com/sl6149/data-scientist-job-market-in-the-us More info available here: https://github.com/Silvialss/projects/tree/master/IndeedWebScraping.
[10] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis *(HLT '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 142–150.
[11] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2019. Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. *IEEE Trans. Pattern Anal. Mach. Intell.* (2019).
[12] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and Optimizing LSTM Language Models. *CoRR* abs/1708.02182 (2017).
[13] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer Sentinel Mixture Models. *CoRR* abs/1609.07843 (2017).
[14] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures *(WSDM '15)*. ACM, New York, NY, USA, 399–408. https://doi.org/10.1145/2684822.2685324
[15] Shirley Eiko Sanematsu. 2001. Taking a broader view of treatment disputes beyond managed care: Are recent legislative efforts the cure? *UCLA Law Review* 48 (2001).
[16] Leslie N. Smith. 2017. Cyclical learning rates for training neural networks. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on. IEEE*. 464–472.
[17] Mark Steyvers and Tom Griffiths. 2007. *Probabilistic Topic Models*. Lawrence Erlbaum Associates.
[18] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Advances in Neural Information Processing Systems*. 3320–3328.
[19] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. *CoRR* abs/1509.01626 (2015). arXiv:1509.01626 http://arxiv.org/abs/1509.01626

Knowledge Intensive Learning of Generative Adversarial Networks

Devendra Singh Dhami
devendra.dhami@utdallas.edu
The University of Texas at Dallas

Mayukh Das
Samsung Research India
mayukh.das@samsung.com

Sriraam Natarajan
The University of Texas at Dallas
sriraam.natarajan@utdallas.edu

ABSTRACT

While Generative Adversarial Networks (GANs) have accelerated the use of generative modelling within the machine learning community, most of the applications of GANs are restricted to images. The use of GANs to generate clinical data has been rare due to the inability of GANs to faithfully capture the intrinsic relationships between features. We hypothesize and verify that this challenge can be mitigated by incorporating domain knowledge in the generative process. Specifically, we propose human-allied GANs that using correlation advice from humans to create synthetic clinical data. Our empirical evaluation demonstrates the superiority of our approach over other GAN models.

CCS CONCEPTS

• **Deep Learning** → **Generative Adversarial Networks**; • **Application** → *Healthcare*; • **Learning** → *Knowledge Intensive Learning*.

KEYWORDS

generative adversarial networks, human in the loop, healthcare

ACM Reference Format:
Devendra Singh Dhami, Mayukh Das, and Sriraam Natarajan. 2020. Knowledge Intensive Learning of Generative Adversarial Networks. In *Proceedings of KDD Workshop on Knowledge-infused Mining and Learning (KiML'20).* , 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Deep learning models have reshaped the machine learning landscape over the past decade [16, 29]. Specifically, Generative Adversarial Networks (GANs) [17] have found tremendous success in generating examples for images [34, 37, 45], photographs of human faces [1, 25, 52], image to image translation [30, 33, 55] and 3D object generation [44, 51, 53] to name a few. Despite such success, there are several key factors that limit the widespread adoption of GANs, for a broader range of tasks, including, widely acknowledged data hungry nature of such methods, potential access issues of real medical data and finally, their restricted usage, mainly in the context of images. These factors have limited the use of these arguably successful techniques in medical (or similar) domains. However, recently, synthetic data generation has become a centerpiece of research in medical AI due to the diverse difficulties in collection, persistence, sharing and analysis of real clinical data.

In M. Gaur, A. Jaimes, F. Ozcan, S. Shah, A. Sheth, B. Srivastava, Proceedings of the Workshop on Knowledge-infused Mining and Learning (KDD-KiML 2020). San Diego, California, USA, August 24, 2020.
KiML'20, August 24, 2020, San Diego, California, USA,

https://doi.org/10.1145/nnnnnnn.nnnnnnn

We aim to address the above limitations. Inspired by Mitchell's argument of "The Need for Biases in Learning Generalizations" [38], we mitigate the challenges of existing data hungry methods via inductive bias while learning GANs. We show that effective inductive bias can be provided by humans in the form of domain knowledge [14, 27, 41, 50]. Rich human advice can effectively balance the impact of quality (sparsity) of training data. Data quality also contributes to, the well studied, modal instability of GANs. This problem is especially critical in domains such as medical/clinical analytics that does not typically exhibit 'spatial homophily' [21], unlike images, and are prone to distributional diversity among feature clusters as well. Our human-guided framework proposes a robust strategy to address this challenge. Note that in our setting the human is an ally and not an adversary.

The second limitation of access is crucial for medical data generation. Access to existing medical databases [10, 18] is hard due to cost and access concerns and thus synthetic data generation holds tremendous promise [6, 13, 19, 35, 48]. While previous methods generated synthetic images, we go beyond images and generate clinical data. Building on this body of work, we present a synthetic data generation framework that effectively exploits domain expertise to handle data quality.

We make a few key contributions:

(1) We demonstrate how effective human advice can be provided to a GAN as an inductive bias.
(2) We present a method for generating data given this advice.
(3) Finally, we demonstrate the effectiveness and efficacy of our approach on 2 de-identified clinical data sets. Our method is generalizable to multiple modalities of data and is not necessarily restricted to images.
(4) Yet another feature of this approach is that training occurs from very few data samples (< 50 in one domain) thus providing human guidance as a data generation alternative.

2 RELATED WORK

The key principle behind GANs [17] is a zero-sum game [26] from game theory, a mathematical representation where each participant's gain or loss is exactly balanced by the losses or gains of the other participants and is generally solved by a minimax algorithm. The generator distribution $p_{data}(\boldsymbol{x})$ over the given data $\boldsymbol{x}$ is learned by sampling $\boldsymbol{z}$ from a random distribution $p_{\boldsymbol{z}}(\boldsymbol{z})$ (initially uniform was proposed but Gaussians have been proven superior [2]). While GANs have proven to be a powerful framework for estimating generative distributions, convergence dynamics of naive mini-max algorithm has been shown to be unstable. Some recent approaches, among many others, augment learning either via statistical relationships between true and learned generative distributions such as Wasserstein-1

distance [3], MMD [32] or via spectral normalization of the parameter space of the generator [39] which controls the generator distribution from drifting too far. Although these approaches have improved the GAN learning in some cases, there is room for improvement.

Guidance via human knowledge is a provably effective way to control learning in presence of systematic noise (which leads to instability). One typical strategy to incorporate such guidance is by providing *rules over training examples and features*. Some of the earliest approaches are explanation-based learning (EBL-NN, [49]) or ANNs augmented with symbolic rules (KBANN, [50]). Various widely-studied techniques of leveraging domain knowledge for optimal model generalization include polyhedral constraints in case of knowledge-based SVMs, [9, 14, 28, 47]), preferences rules [5, 27, 41, 42] or qualitative constraints (ex: monotonicities / synergies [54] or quantitative relationships [15]). Notably, whereas these models exhibit considerable improvement with the incorporation of human knowledge, there is only limited use of such knowledge in training GANs. Our approach resembles the qualitative constraints framework in spirit.

While widely successful in building optimally generalized models in presence of systematic noise (or sample biases), knowledge-based approaches have mostly been explored in the context of discriminative modeling. In the generative setting, a recent work extends the principle of posterior regularization from Bayesian modeling to deep generative models in order to incorporate structured domain knowledge [22]. Traditionally, knowledge based generative learning has been studied as a part of learning probabilistic graphical models with structure/parameter priors [36]. We aim to extend the use of knowledge to the generative model setting.

3 KNOWLEDGE INTENSIVE LEARNING OF GENERATIVE ADVERSARIAL NETWORKS

A notable disadvantage of adversarial training formulation is that the training is slow and unstable, leading to mode collapse [2] where the generator starts generating data of only a single modality. This has resulted in GANs not being exploited to their full potential in generating synthetic non-image clinical data. Human advice can encourage exploration in diverse areas of the feature space and helps learn more stable models [43]. Hence, we propose a human-allied GAN architecture (HA-GAN) (figure 1). The architecture incorporates human advice in form of feature correlations. Such intrinsic relationships between the features are crucial in medical data sets and thus become a natural candidate as additional knowledge/advice in guided model learning for faithful data generation.

Our approach builds upon a GAN architecture [17] where a random noise vector is provided to the generator which tries to generate examples as close to the real distribution as possible. The discriminator tries to distinguish between real examples and ones generated by the generator. The generator tries to maximize the probability that the discriminator makes a mistake and the discriminator tries to minimize its mistakes thereby resulting in a min-max optimization problem which can be solved by a mini-max algorithm. We adopt the Wasserstein GAN (WGAN) architecture[1] [3, 20] that focuses on defining a distance/divergence (Wasserstein or earth movers distance) to measure the closeness between the real distribution and the model distribution.

[1]We use 'GAN' to indicate 'W-GAN'

3.1 Human input as inductive bias

Historically, two approaches have been studied for using guidance as bias. The **first** is to provide advice on the labels as constraints or preferences that controls the search space. Some example advice rules on the labels include: *(3 ≤ feature$_1$ ≤ 5) ⇒ label = 1* and *(0.6 ≤ feature$_2$ ≤ 0.8) ∧ (4 ≤ feature$_3$ ≤ 5) ⇒ label = 0*. Such advice is more relevant in an discriminative setting but are not ideal for GANs. Since GANs are shown to be sensitive to the training data and here the *labels are getting generated*, they should not be altered during training. The **second** is via correlations between features as preferences (*our approach*) which allows for faithful representation of diverse modality.

Advice injection: After every fixed number of iterations, N, we calculate the correlation matrix of the generated data $\mathcal{G}_1$ and provide a set of advice ψ on the correlations between different features. Consider the following motivating example for the use of correlations as a form of advice.

Example: *Consider predicting heart attack with 3 features - cholesterol, blood pressure (BP) and income. The values of the given features can vary (sometimes widely) between different patients due to several latent factors (ex, smoking habits). It is difficult to assume any specific distribution. In other words, it is difficult to deduce whether the values for the features come from the same distribution (even though the feature values in the data set are similar).*

We modify the correlation coefficients (for both positive and negative correlations) between the features by increasing them if the human advice suggests that two features are highly correlated and decrease the same if the advice suggests otherwise.

Example: *Continuing the above example, since rise in the cholesterol level can lead to rise in BP and vice versa, expert advice here can suggest that cholesterol and BP should be highly correlated. Also, as income may not contribute directly to BP and cholesterol levels, another advice here can be to de-correlate cholesterol/BP and income level.*

The example advice rules ∈ ψ are: 1. Correlation("cholesterol level","BP")↑, 2. Correlation("cholesterol level","income level")↓ and 3. Correlation("BP","income level")↓, where ↑ and ↓ indicate increase and decrease respectively. Based on the 1st advice we need to increase the correlation coefficient between *cholesterol level* and *BP*. Then

$$C = \begin{bmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix} \mathcal{A} = \begin{bmatrix} 1 & \lambda & 1 \\ \lambda & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad (1)$$

Here C is the correlation matrix, $\mathcal{A}$ is the advice matrix and λ is the factor by which the correlation value is to be augmented. In case where we need to increase the value of the correlation coefficient, λ should be > 1. We keep $\lambda = \frac{1}{max(|C|)}$. Since -1.0 ≤ $\forall c \in C \leq 1.0$, in this case, the value of $\lambda \geq 1.0$, leading to enhanced correlation via

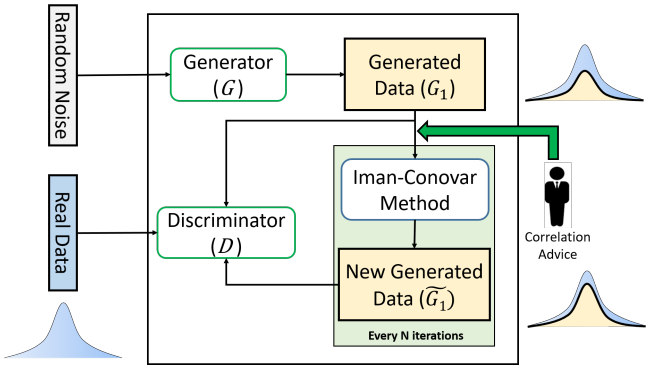


Figure 1: Human-Allied GAN. Correlation advice takes generated distribution closer to the real distribution.

Hadamard product. Thus the new correlation matrix $\hat{C}$ is,

$$\hat{C} = C \odot \mathcal{A} = \begin{bmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix} \odot \begin{bmatrix} 1 & \frac{1}{0.3} & 1 \\ \frac{1}{0.3} & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.667 & 0.3 \\ 0.667 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix} \tag{2}$$

If the advice says that features have low correlations (2nd rule in example), we decrease the correlation coefficient. Now, λ must be < 1 and we set $\lambda = max(|C|)$. Since -1 $\leq \forall c \in C \leq 1.0$, the value of $\lambda \leq 1.0$. Thus multiplying by λ will decrease the correlation value, and the new correlation matrix is,

$$\hat{C}_1 = \hat{C} \odot \mathcal{A} = \begin{bmatrix} 1 & 0.667 & 0.3 \\ 0.667 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix} \odot \begin{bmatrix} 1 & 1 & 0.3 \\ 1 & 1 & 0.3 \\ 0.3 & 0.3 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.667 & 0.09 \\ 0.667 & 1 & 0.021 \\ 0.09 & 0.021 & 1 \end{bmatrix} \tag{3}$$

This is used to create the new generated data $\tilde{\mathcal{G}}_1$. For negative correlations, the process is unchanged.

3.2 Advice-guided data generation

After $\hat{C}_1$ is constructed, we next generate data satisfying the constraints. To this effect, we employ the Iman-Conover method [23], a distribution free method to define dependencies between distributional variables based on rank correlations such as Spearman or Kendell Tau correlations. Since we deal with linear relationships between the features and assume a normal distribution and that Pearson coefficient has shown to perform equally well with the Iman-Conover method [40] due to the close relationship between Pearson and Spearman correlations, we use the Pearson correlations. Further, we assume that the features are Gaussian, justified by the fact that most lab test data is continuous. The Iman-Conover method consists of the following steps:

[Step 1]: Create a random standardized matrix $\mathcal{M}$ with values $x \in \mathcal{M} \sim$ Gaussian distribution. This is obtained by the process of inverse transform sampling described next. Let $\mathcal{V}_1$ be a uniformly distributed random variable and $\mathcal{CDF}$ be the cumulative distribution function. For a sampled point v, $\mathcal{CDF}(v) = \mathcal{P}(V \leq v)$. Thus, to generate samples, the values $v \sim \mathcal{V}$ are passed through $\mathcal{CDF}^{-1}$ to obtain the desired values x $[\mathcal{CDF}^{-1}(v) = \{x | \mathcal{CDF}(x) \leq v, v \in [0,1]\}]$. Thus for Gaussian,

$$\mathcal{CDF}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp^{\frac{-x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{0}^{x} \exp^{\frac{-x^2}{2}} dx = [-\exp(\frac{-x^2}{2})]_0^x \tag{4}$$

The inverse CDF can be thus written as $\mathcal{CDF}^{-1}(v) = 1 - \exp(\frac{-x^2}{2}) \leq v$ and the desired values $x \in \mathcal{M}$ can be obtained as $x = \sqrt{2ln(1-v)}$.

[Step 2]: Calculate the correlation matrix $\mathcal{E}$ of $\mathcal{M}$.

[Step 3]: Calculate the Cholesky decomposition $\mathcal{F}$ of the correlation matrix $\mathcal{E}$. Cholesky decomposition [46] of a *positive-definite* matrix is given as the product of a lower triangular matrix and its conjugate transpose. Note that for Cholesky decomposition to be unique, the target matrix should be positive definite, (such as the co-variance matrix) whereas the correlation matrix, used in our algorithm, is only positive semi-definite. We enforce positive-definiteness by repeated addition of very small values to the diagonal of the correlation matrix until positive-definiteness is ensured. Given a symmetric and positive definite matrix $\mathcal{E}$, its Cholesky decomposition $\mathcal{F}$ is such that $\mathcal{E} = \mathcal{F} \cdot \mathcal{F}^\top$.

[Step 4]: Calculate the Cholesky decomposition $\mathcal{Q}$ of the correlation matrix obtained after modifications based on human advice, $\hat{C}$. As above the Cholesky decomposition is such that $\hat{C} = \mathcal{Q} \cdot \mathcal{Q}^\top$.

[Step 5]: Calculate the reference matrix $\mathcal{T}$ by transforming the sampled matrix $\mathcal{M}$ from step 1 to have the desired correlations of $\hat{C}$, by using their Cholesky decompositions.

[Step 6]: Rearrange values in columns of the generated data $\mathcal{G}_1$ to have the same ordering as corrresponding column in the reference matrix $\mathcal{T}$ to obtain the final generated data $\tilde{\mathcal{G}}_1$.

Cholesky decomposition to model correlations: Given an randomly generated data set with no correlations $\mathcal{P}$, a correlation matrix C and its Cholesky decomposition $\mathcal{Q}$, data that faithfully follows the given correlations $\in C$ can be generated by the product of the obtained lower triangular matrix with the original uncorrelated data

i.e. $\hat{\mathcal{P}}$=$Q\mathcal{P}$. The correlation of the newly obtained data, $\hat{\mathcal{P}}$ is,

$$Corr(\hat{\mathcal{P}}) = \frac{Cov(\hat{\mathcal{P}})}{\sigma_{\hat{\mathcal{P}}}} = \frac{\mathbf{E}[\hat{\mathcal{P}}\hat{\mathcal{P}}^\top] - \mathbf{E}[\hat{\mathcal{P}}]\mathbf{E}[\hat{\mathcal{P}}]^\top}{\sigma_{\hat{\mathcal{P}}}} \tag{5}$$

Since we consider data $\hat{\mathcal{P}}$ from a Gaussian distribution with zero mean and unit variance,

$$\begin{aligned} Corr(\hat{\mathcal{P}}) &= \frac{\mathbf{E}[\hat{\mathcal{P}}\hat{\mathcal{P}}^\top] - \mathbf{E}[\hat{\mathcal{P}}]\mathbf{E}[\hat{\mathcal{P}}]^\top}{\sigma_{\hat{\mathcal{P}}}} = \mathbf{E}[\hat{\mathcal{P}}\hat{\mathcal{P}}^\top] = \mathbf{E}[(Q\mathcal{P})(Q\mathcal{P})^\top] \\ &= \mathbf{E}[Q\mathcal{P}Q^\top\mathcal{P}^\top] = Q\mathbf{E}[\mathcal{P}\mathcal{P}^\top]Q^\top = QQ^\top = C \end{aligned} \tag{6}$$

Thus Cholesky decomposition can capture the desired correlations faithfully and can be used for generating correlated data. Since we already have a normal sampled matrix $\mathcal{M}$ and a calculated correlation $\mathcal{E}$ of $\mathcal{M}$, we need to calculate a reference matrix (step 5).

3.3 Human-Allied GAN training

Since the human expert advice is provided independent of the GAN architecture, our method is agnostic of the underlying GAN architecture. We make use of Wasserstein GAN (WGAN) architecture since its shown to be more stable while training and can handle mode collapse [3]. Only the error backpropagation values differ when we are using the data generated by the underlying GAN or the data generated by the Iman-Conover method. Our algorithm starts with the general process of training a GAN where the generator takes random noise as an input and generates data which is then passed, along with the real data, to the discriminator. The discriminator tries to identify the real and generated data and the error is back propagated to the generator. After every specified number of iterations, the correlations between features C in the generated data is obtained and a new correlation matrix $\hat{C}$, is obtained with respect to the expert advice (section 3.1). A new data set is generated wrt $\hat{C}$ using the Iman-Conover method (Section 3.2) and then passed to the discriminator along with the real data set.

4 EXPERIMENTAL EVALUATION

We aim to answer the following questions:

- **Q1:** Does providing advice to GANs help in generating better quality data?
- **Q2:** Are GANs with advice effective for data sets that have few examples?
- **Q3:** How does bad advice affect the quality of generated data?
- **Q4:** How well does human advice handle class imbalance?
- **Q5:** How does our method compare to state-of-the-art GAN architectures.

We consider 2 *real clinical data sets*.

(1) **Nephrotic Syndrome** is a novel data set of symptoms that indicate kidney damage. This consists of 50 kidney biopsy images along with the clinical reports sourced from Dr Lal PathLabs, India [2]. We use the clinical reports that consist of the values for kidney tissue diagnosis which can confirm the clinical diagnosis and help to identify high-risk patients and influence treatment decisions and help medical practitioners to plan and prognosticate treatments. The data consists of 19 features with 44 positive and 6 negative examples.

(2) **MIMIC** database [24] consists of deidentified information of patients admitted to critical care units at a large tertiary care hospital. The features included are predominately time window aggregations of physiological measurements from the medical records. We selected relevant lab results, vital sign observations and feature aggregations. The data consists of 18 with 5813 positive and 40707 negative examples.

Advice Acquisition: Here we compile the sources from which we obtain the advice.

(1) *Nephrotic Syndrome:* This is a novel real data set and the **advice is obtained from a nephrologist in India**. According to the problem statement from the expert, nephrotic syndrome involves the loss of a lot of protein and nephritic syndrome involves the loss of a lot of blood through urine. A kidney biopsy is often required to diagnose the underlying pathology in patients with suspected glomerular disease. The goal of the project is to build a clinical support system that predicts the disease using clinical features, thus reducing the need of kidney biopsy. ***Since the data collection is scarce, a synthetic data set can help in better understanding of the disease from the clinical features.***

(2) *MIMIC:* The feature set and the expected correlations are obtained in consultation with **trauma experts at a Dallas hospital**.

All experiments were run on a *64-bit Intel(R) Xeon(R) CPU E5-2630 v3* server for 10K epochs. Both the generator and discriminator are neural networks with 4 hidden layers. To measure the quality of the generated data we make use of the *train on synthetic, test on real* (TSTR) method as proposed in [12]. We use gradient boosting with 100 estimators and a learning rate of 0.01 as the underlying model. We train the GAN for 10K epochs and provide correlation advice every 1K iterations.

Table 1 shows the results of the TSTR method with data generated with (HA-GAN$_{GA}$) and without advice (GAN). It shows that the data generated with advice has higher TSTR performance than the data generated without advice across all data sets and all metrics. Thus, to answer **Q1**, providing advice to generative adversarial networks captures the relationship between features better and thus are able to generate better quality synthetic data.

Learning with less data: GANs with advice are especially impressive in nephrotic syndrome data which consists of only 50 examples across all metrics and is thus very small in size when compared to the number of samples typically required to train a GAN model. Thus, we realize an important property of incorporating human guidance in the GAN model and can answer **Q2** affirmatively. *The use of advice opens up the potential of using GANs in presence of sparse data samples*.

Effect of bad advice: Table 1 also shows the results for data generated with bad advice (HA-GAN$_{BA}$). To simulate bad advice, we follow a simple process: if the advice says that the correlation between features should be high, we set the correlations in $\hat{C}$ to 0 and if the advice says that the correlation should be low, we set the correlations in $\hat{C}$ to be either 1 or -1 based on whether the original

[2] https://www.lalpathlabs.com/

Table 1: TSTR Results ($\approx$ 3 *dec.*). N/A in Nephrotic Syndrome denotes that all generated labels were of a single class (0 in our case) and thus we were not able to run the discriminative algorithm in the TSTR method. *GA* and *BA* denotes good and bad advice to our HA-GAN model respectively.

Data set	Methods	Recall	F1	AUC-ROC	AUC-PR
NS	GAN	0.584	0.666	0.509	0.911
	HA-GAN$_{BA}$	0.42	0.511	0.518	0.886
	medGAN	N/A	N/A	N/A	N/A
	medWGAN	N/A	N/A	N/A	N/A
	medBGAN	N/A	N/A	N/A	N/A
	HA-GAN$_{GA}$	**1.0**	**0.943**	**0.566**	**0.947**
MIMIC	GAN	0.122	0.119	0.495	0.174
	HA-GAN$_{BA}$	0.285	0.143	0.459	0.235
	medGAN	0.374	0.163	0.478	0.279
	medWGAN	0.0	0.0	0.5	0.562
	medBGAN	0.0	0.0	0.5	0.562
	HA-GAN$_{GA}$	**0.979**	**0.263**	**0.598**	**0.567**

correlation is positive or negative. Thus, given a correlation matrix

$$C = \begin{bmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix} \tag{7}$$

suppose the advice says that we need to increase the correlation coefficient between feature 1 and feature 2. Then the new correlation matrix after bad advice can be calculated as:

$$C = \begin{bmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix} \mathcal{A} = \begin{bmatrix} 1 & \lambda & 1 \\ \lambda & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{8}$$

$$\hat{C} = C \odot \mathcal{A} = \begin{bmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix} \odot \begin{bmatrix} 1 & \lambda & 1 \\ \lambda & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{9}$$

where λ is the factor by which the correlation value is to be augmented. Since the advice asks to increase the correlation, we set λ=0. Thus,

$$\hat{C} = \begin{bmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix} \odot \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.0 & 0.3 \\ 0.0 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix} \tag{10}$$

Similarly, if the advice says that we need to decrease the correlation coefficient between feature 1 and feature 3, we set $\lambda = \frac{1}{feat_{val}}$.

$$\hat{C} = \begin{bmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix} \odot \begin{bmatrix} 1 & 0.2 & \frac{1}{0.3} \\ 0.2 & 1 & 1 \\ \frac{1}{0.3} & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.2 & 1.0 \\ 0.2 & 1 & 0.07 \\ 1.0 & 0.07 & 1 \end{bmatrix} \tag{11}$$

As results show in table 1, giving bad advice adversely affects the performance thereby answering **Q3**.

The nephrotic syndrome and MIMIC data sets are relatively unbalanced with a pos to neg ratio of $\approx$ 8:1 and 1:7 respectively. Most of the medical data sets, except highly curated data sets, are unbalanced. A data generator model should be able to handle this imbalance. Since our method explicitly focuses on the correlations between features and generates better quality data based on such relationships between features, our method is quite **robust to the imbalance in the underlying data**. This can be seen in the results in table 1 where advice based data generation outperforms the non-advice and bad advice based data generation. Thus, we can answer **Q4** affirmatively.

To answer **Q5** we compare our method to 3 GAN architectures, medGAN [8] which uses an encoder decoder framework for EHR data generation and its 2 variants medBGAN and medWGAN [4] and the results are shown in table 1. Our method, with good advice, outperforms the baseline both domains showing the effectiveness of our method.

5 CONCLUSION

We presented a new GAN formulation that employs correlation information between features as advice to generate new correlated data and train the underlying GAN model. We tested our model on real clinical data sets and show that incorporating advice helps generate good quality synthetic medical data. We employ TSTR method to test the quality of generated data and demonstrated that the generated data with advice is more aligned with the real data. There are several future interesting directions. First, providing advice only when required in an active fashion can allow for significant reduction in the amount of effort on the human side. Second, there can be multiple advice options, such as posterior regularization [15], that can be used to capture feature relationships explicitly. Third, although we do not have identifiers in the data, thereby eliminating the need of differential privacy [11], a general framework that can uphold the privacy of patient data along the lines of using Cholesky decomposition [7, 31] is a natural next step.

ACKNOWLEDGMENTS

DSD and SN gratefully acknowledge DARPA Minerva award FA9550-19-1-0391. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the DARPA or the US government.

REFERENCES

[1] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. 2017. Face aging with conditional generative adversarial networks. In *ICIP*.
[2] Martin Arjovsky and Leon Bottou. 2017. Towards principled methods for training generative adversarial networks. In *ICLR*.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. *ICML* (2017).
[4] Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. 2019. Synthesizing electronic health records using improved generative adversarial networks. *JAMA* (2019).
[5] Darius Braziunas and Craig Boutilier. 2006. Preference elicitation and generalized additive utility. In *AAAI*.
[6] Anna L Buczak, Steven Babin, and Linda Moniz. 2010. Data-driven approach for creating synthetic electronic medical records. *BMC medical informatics and decision making* (2010).
[7] Jim Burridge. 2003. Information preserving statistical obfuscation. *Statistics and Computing* (2003).
[8] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. 2017. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. In *MLHC*.
[9] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* (1995).
[10] Ivo D Dinov. 2016. Volume and value of big healthcare data. *Journal of medical statistics and informatics* (2016).
[11] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *TAMS*.
[12] Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. 2017. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633* (2017).
[13] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. 2018. Synthetic data augmentation using GAN for improved liver lesion classification. In *ISBI*.
[14] Glenn M Fung, Olvi L Mangasarian, and Jude W Shavlik. 2003. Knowledge-based support vector machine classifiers. In *NIPS*.
[15] Kuzman Ganchev, Jennifer Gillenwater, Ben Taskar, et al. 2010. Posterior regularization for structured latent variable models. *JMLR* (2010).
[16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*.
[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*.
[18] Peter Groves, Basel Kayyali, David Knott, and Steve Van Kuiken. 2016. The'big data'revolution in healthcare: Accelerating value and innovation. (2016).
[19] John T Guibas, Tejpal S Virdi, and Peter S Li. 2017. Synthetic medical images from dual generative adversarial networks. *arXiv preprint arXiv:1709.01872* (2017).
[20] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *NIPS*.
[21] Haroun Habeeb, Ankit Anand, Mausam Mausam, and Parag Singla. 2017. Coarse-to-fine lifted MAP inference in computer vision. In *IJCAI*.
[22] Zhiting Hu, Zichao Yang, Russ R Salakhutdinov, LIANHUI Qin, Xiaodan Liang, Haoye Dong, and Eric P Xing. 2018. Deep Generative Models with Learnable Knowledge Constraints. In *NeurIPS*.
[23] Ronald L Iman and William-Jay Conover. 1982. A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics-Simulation and Computation* (1982).
[24] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* (2016).
[25] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*.
[26] Harold William Kuhn and Albert William Tucker. 1953. *Contributions to the Theory of Games*.
[27] Gautam Kunapuli, Phillip Odom, Jude W Shavlik, and Sriraam Natarajan. 2013. Guiding autonomous agents to better behaviors through human advice. In *ICDM*.
[28] Quoc V Le, Alex J Smola, and Thomas Gärtner. 2006. Simpler knowledge-based support vector machines. In *ICML*.
[29] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* (2015).
[30] Minjun Li, Haozhi Huang, Lin Ma, Wei Liu, Tong Zhang, and Yugang Jiang. 2018. Unsupervised image-to-image translation with stacked cycle-consistent adversarial networks. In *ECCV*.
[31] Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang. 2011. Enabling multilevel trust in privacy preserving data mining. *TKDE* (2011).
[32] Yujia Li, Kevin Swersky, and Rich Zemel. 2015. Generative moment matching networks. In *ICML*.
[33] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. In *NIPS*.
[34] Ming-Yu Liu and Oncel Tuzel. 2016. Coupled generative adversarial networks. In *NIPS*.
[35] Faisal Mahmood, Richard Chen, and Nicholas J Durr. 2018. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. *IEEE transactions on medical imaging* (2018).
[36] V. K. Mansinghka, C. Kemp, J. B. Tenenbaum, and T. L. Griffiths. 2006. Structured Priors for Structure Learning. In *UAI*.
[37] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *ICCV*.
[38] Tom M Mitchell. 1980. *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ. New Jersey.
[39] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral normalization for generative adversarial networks. *ICLR* (2018).
[40] Klemen Naveršnik and Klemen Rojnik. 2012. Handling input correlations in pharmacoeconomic models. *Value in Health* (2012).
[41] P. Odom, T. Khot, R. Porter, and S. Natarajan. 2015. Knowledge-Based Probabilistic Logic Learning. In *AAAI*.
[42] Phillip Odom and Sriraam Natarajan. 2015. Active advice seeking for inverse reinforcement learning. In *AAAI*.
[43] Phillip Odom and Sriraam Natarajan. 2018. Human-guided learning for probabilistic logic models. *Frontiers in Robotics and AI* (2018).
[44] Michela Paganini, Luke de Oliveira, and Benjamin Nachman. 2018. CaloGAN: Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks. *Physical Review D* (2018).
[45] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR* (2016).
[46] Ernest M Scheuer and David S Stoller. 1962. On the generation of normal random vectors. *Technometrics* (1962).
[47] Bernhard Schölkopf, Patrice Simard, Alex J Smola, and Vladimir Vapnik. 1998. Prior knowledge in support vector kernels. In *Advances in neural information processing systems*. 640–646.
[48] Rittika Shamsuddin, Barbara M Maweu, Ming Li, and Balakrishnan Prabhakaran. 2018. Virtual patient model: an approach for generating synthetic healthcare time series data. In *ICHI*.
[49] Jude W Shavlik and Geoffrey G Towell. 1989. Combining explanation-based learning and artificial neural networks. In *Proceedings of the sixth international workshop on Machine learning*. Elsevier.
[50] Geoffrey G Towell and Jude W Shavlik. 1994. Knowledge-based artificial neural networks. *Artificial intelligence* (1994).
[51] Yan Wang, Biting Yu, Lei Wang, Chen Zu, David S Lalush, Weili Lin, Xi Wu, Jiliu Zhou, Dinggang Shen, and Luping Zhou. 2018. 3D conditional generative adversarial networks for high-quality PET image estimation at low dose. *NeuroImage* (2018).
[52] Zongwei Wang, Xu Tang, Weixin Luo, and Shenghua Gao. 2018. Face aging with identity-preserved conditional generative adversarial networks. In *CVPR*.
[53] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. 2016. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NIPS*.
[54] S. Yang and S. Natarajan. 2013. Knowledge Intensive Learning: Combining Qualitative Constraints with Causal Independence for Parameter Learning in Probabilistic Models. In *ECMLPKDD*.
[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.

Depressive, Drug Abusive, or Informative: Knowledge-aware Study of News Exposure during COVID-19 Outbreak

Amanuel Alambo
Knoesis Center
Dayton, Ohio
amanuel@knoesis.org

Manas Gaur
AI Institute, University of South Carolina
Columbia, South Carolina
mgaur@email.sc.edu

Krishnaprasad Thirunarayan
Knoesis Center
Dayton, Ohio
tkprasad@knoesis.org

ABSTRACT

The COVID-19 pandemic is having a serious adverse impact on the lives of people across the world. COVID-19 has exacerbated community-wide depression, and has led to increased drug abuse brought about by isolation of individuals as a result of lockdown. Further, apart from providing informative content to the public, the incessant media coverage of COVID-19 crisis in terms of news broadcasts, published articles and sharing of information on social media have had the undesired snowballing effect on stress levels (further elevating depression and drug use) due to uncertain future. In this position paper, we propose a novel framework for assessing the spatio-temporal-thematic progression of depression, drug abuse, and informativeness of the underlying news content across the different states in the United States. Our framework employs an attention-based transfer learning technique to apply knowledge learned on a social media domain to a target domain of media exposure. To extract news articles that are related to COVID-19 communications from the streaming news content on the web, we use neural semantic parsing, and background knowledge bases in a sequence of steps called semantic filtering. We achieve promising preliminary results on three variations of Bidirectional Encoder Representations from Transformers (BERT) model. We compare our findings against a report from Mental Health America and the results show that our fine-tuned BERT models perform better than vanilla BERT. Our study can benefit epidemiologists by offering actionable insights on COVID-19 and its regional impact. Further, our solution can be integrated into end-user applications to tailor news for users based on their emotional tone measured on the scale of depressiveness, drug abusiveness, and informativeness.

KEYWORDS

COVID-19; Spatio-Temporal-Thematic; Depressiveness; Drug Abuse; Informativeness; Transfer Learning

ACM Reference Format:
Amanuel Alambo, Manas Gaur, and Krishnaprasad Thirunarayan. 2020. Depressive, Drug Abusive, or Informative: Knowledge-aware Study of News Exposure during COVID-19 Outbreak . In *Proceedings of KDD Workshop on Knowledge-infused Mining and Learning (KiML'20).* , 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

In M. Gaur, A. Jaimes, F. Ozcan, S. Shah, A. Sheth, B. Srivastava, Proceedings of the Workshop on Knowledge-infused Mining and Learning (KDD-KiML 2020). San Diego, California, USA, August 24, 2020.
KiML'20, August 24, 2020, San Diego, California, USA,

https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

COVID-19 pandemic has changed our societal dynamics in different ways due to the varying impact of the news articles and broadcasts on a diverse population in the society. Thus, it is important to place the news articles in their spatio-temporal-thematic (Nagarajan et al., 2009; Andrienko et al., 2013; Harbelot et al., 2015) contexts to offer appropriate and timely response and intervention. In order to limit the scope of this research agenda, we propose to focus on identifying regions that are exposed to depressive and drug abusive news articles and to determine/recommend ways for timely interventions by epidemiologists.

The impact of COVID-19 on mental health has been investigated in recent studies (Garfin et al., 2020; Holmes et al., 2020; Qiu et al., 2020). [4] studied the impact of repeated media exposure on the mental well-being of individuals and its ripple effects. [8] underscore the importance of a multidisciplinary study to better understand COVID-19. Specifically, the study explores its psychological, social, and neuroscientific impacts. [12] studied the psychological impact COVID-19 lockdown had on the Chinese population. These studies, however, do not adequately explore a technique to computationally analyze the regional repercussions associated with media exposure to COVID-19 that may provide a better basis for local grassroots level action.

We propose an approach to measure depressiveness, drug abusiveness, and informativeness as a result of media exposure for various states in the US in the months from January 2020 to March 2020. Our study is focused on the first quarter of 2020 as this period was critical in the spread of COVID-19 and its ominous impact; this was a period when the public faced major changes to lifestyle including lockdown, social distancing, closure of businesses, unemployment, and broadly speaking, complete lack of control over the unfolding situation precipitating in severe uncertainty about the impending future. In consequence, this continued media exposure progressively worsened the mental health of individuals across the board. We analyze and score news content on three orthogonal dimensions: spatial, temporal, and thematic. For spatial, we use state boundaries. For temporal, we use monthly data analysis. For thematic, we score news content on the category/dimension of depression, drug abuse and informativeness (relevant to COVID-19 but not directly connected to either depression or drug-abuse).

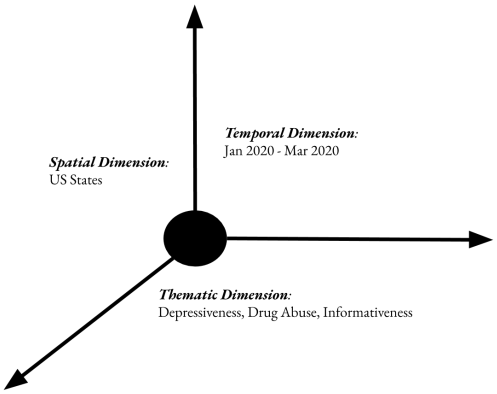


Figure 1: Spatio-Temporal-Thematic Dimensions

Our study hinges on the use of domain-specific language modeling and transfer learning to better understand how depressiveness, drug abusiveness, and informativeness of news articles evolve in response to media exposure by people. We conduct the transfer of knowledge learned on a social media platform to the domain of exposure to news using variations of the attention-based BERT model (Devlin et al., 2018), also called Vanilla BERT. Thus, in addition to vanilla BERT, we fine-tune BERT models on corpora that are representative of depression and drug abuse. Then, we compare results obtained using the three variants of the BERT model. For scoring depressiveness, drug abusiveness, and informativeness of news articles, we utilize entities from structured domain knowledge from the Patient Health Questionnaire (PHQ-9) lexicon (Yazdavar et al., 2017), Drug Abuse Ontology (DAO) (Cameron et al., 2013), and DBpedia (Lehmann et al., 2015). PHQ-9 lexicon is a knowledge base developed specifically for assessing depression, and DAO is built to study drug abuse. Similarly, we use DBpedia, which is a generic and comprehensive knowledge base, for assessing the informativeness of news content.

Having determined the scores for depressiveness, drug abusiveness, and informativeness of news articles for each state during the three months, we computed the aggregate score for each thematic category by summing up the scores for the news articles. We finally assigned the category with the highest score as a label for a state. For instance, if the aggregate score of depressiveness for the state of Iowa in the month of January 2020 is the highest of the three thematic categories, then the state of Iowa is assigned a label of depression for that month, which means the state of Iowa is most exposed to depressive news contents. Thus, identifying which states are consistently exposed to depressive or drug abusive news contents enables policy makers and epidemiologists to devise appropriate intervention strategies.

2 DATA COLLECTION

We collected 1.2 Million news articles from the Web and GDELT[1] (a resource that stores world news on significant events from different countries) using semantic filtering (Sheth and Kapanipathi, 2016) and spanning the period from January 01, 2020, to March 29, 2020. We filtered news articles that did not originate from within the US and grouped the ones that are from the US based on their state of origination. The state-level grouped news articles had a total of over 150K entities identified using DBpedia spotlight service[2]. However, since using a coarse filtering service such as DBpedia spotlight over the entire news articles is not efficient and brings in irrelevant entities, and thus noisy news articles, we utilize ("i") a neural parsing approach with self-attention (Wu et al., 2019) to extract relevant entities. After extracting relevant entities and news articles, we use ("ii") DBpedia spotlight service to identify news articles that are related to online communications about COVID-19.

Figure 2: Knowledge-based entity extraction using Semantic Filtering

For this task, we explored 780 DBpedia categories that are relevant to COVID-19 communications to create the most relevant set of entities and news articles. Further, upon inspection of the news articles, we discovered medical terms that were not available in DBpedia. As a result, we used ("iii") the MeSH terms hierarchy in Unified Medical Language System (UMLS), the Diagnostic and Statistical Manual for Mental Disorders (DSM-5) lexicon (Gaur et al., 2018), and Drug Abuse Ontology (DAO), collectively referred to as *Mental Health and Drug Abuse Knowledgebase* (MHDA-Kb) to spot additional entities. Thus, from 700K unique news articles (which are extracted from the total of 1.2 Million news articles by removing duplicates), we created a set of 120K unique entities that are described by the 780 DBpedia categories and 225 concepts in MHDA-Kb. The figures below show two examples that illustrate entities spotted during entity extraction on a sample news article. A news article that has entities identified using this sequence of steps is selected for our study.

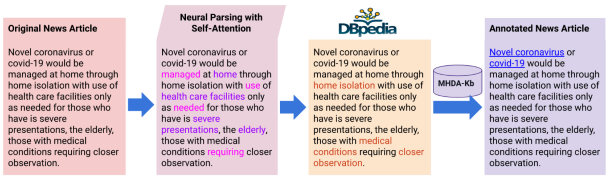


Figure 3: Example entity extraction-I using Semantic Filtering

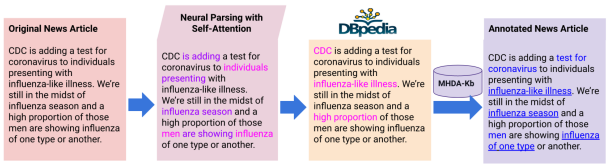


Figure 4: Example entity extraction-II using Semantic Filtering

[1]https://www.gdeltproject.org/

[2]https://www.dbpedia-spotlight.org/

3 METHODS

We propose to use three variations of the BERT model for representing news articles. In its basic form, we use vanilla BERT for encoding news articles. For the remaining two variations, we fine-tune BERT on a binary sequence classification task by independently training on two corpora using masked language modeling (MLM) and next sentence prediction (NSP) objectives. The two corpora used are: 1) Subreddit Depression (Gkotsis et al., 2017; Gaur et al., 2018); 2) A combination of subreddits: Crippling Alcoholism, Opiates, Opiates Recovery, and Addiction (abbreviated COOA), each consisting of Reddit posts about drug abuse. Subreddit Depression has 760049 posts across 121795 Redditors, and COOA has 1416765 posts from 46183 users, both consisting of posts from the years 2005 - 2016. Reddit posts belonging to subreddits depression or COOA are considered positive classes and the 380444 posts from control group (~10K subreddits unrelated to mental health) as negative classes. We use the following settings for training our BERT model for sequence classification: training batch size of 16, maximum sequence length of 256, Adam optimizer with learning rate of 2e-5, number of training epochs set to 10, and a warmup proportion of 0.1. We used 40%-60% split for training and testing sets for creating the BERT models and achieved a test accuracy of 89% for Depression-BERT and 78% for Drug Abuse-BERT. We set the size of the training set smaller than the testing set for generalizability of our models. In this manuscript, we refer to the BERT model fine tuned on subreddit depression as Depression-BERT or DPR-BERT, while the one fine tuned on subreddit COOA as Drug Abuse-BERT or DA-BERT.

In addition to using BERT for encoding news contents, we also use it for representing the entities in the background knowledge bases (i.e., PHQ-9, DAO, and DBpedia). Once we have encoded the news articles and the entities in the knowledge bases using vanilla BERT or fine-tuned BERT model, we generated depressiveness score, drug abusiveness score, and informativeness score corresponding to the entities in PHQ-9, DAO, and DBpedia respectively. The equation below gives the score of a news article for a category given one of the BERT models:

$$Score_c^m(news) = \frac{1}{|\mathrm{E}_{KB}|} \sum_{e=1}^{|\mathrm{E}_{KB}|} cossim\ (\mathrm{news}, e) \qquad (1)$$

where,

m ∈ {vanilla-BERT, DPR-BERT, DA-BERT}

c ∈ {informativeness, depressiveness, drug abuse}

cossim (news, e): cosine similarity between a news content and an entity in KB

KB - a collection of entities present in PHQ-9, DBpedia, or DAO

We used the base variant of the BERT model with 12 layers, 768 hidden units, and 12 attention heads. We use PyTorch 1.5.0+cu101 for fine-tuning our BERT models. All our programs were run on Google Colab's NVIDIA Tesla P100 PCI-E GPU.

4 PRELIMINARY RESULTS AND DISCUSSION

In this section, we report the state-wise labels (i.e., depressive, drug abusive, informative) for each month obtained after summing the scores of news articles as described. The category with the highest cumulative score is set as the label for a state.

Using vanilla-BERT (Figure 5), we can see that no state shows exposure to news content on drug abuse in January. Going from February to March, we see depressive news content move from inner-most states such as Missouri, Kansas, and Colorado to border states such as California, Montana, North Dakota, and Louisiana, making way for informative news content. Further, there are fewer states exposed to drug-related news content than those exposed to depressive or informative news content in February or March. Particularly, Arizona and Virginia show consistent exposure to drug-related news content in February and March.

Using depression-BERT, as shown in Figure 6, we see that states such as Texas, and Kansas are exposed to depressive news content for the month of January and February while states such as California, Montana, Alaska, and Michigan show higher consumption of depressive news content in February and March. With regard to informativeness, we see an overall even distribution of informative news content across the nation in February and March. Further, we see a few midwest states showing relatively higher instances of news content that are informative than depressive in February and March. It's interesting to see a few southern states such as Oklahoma, Texas, and Arkansas transition from exposure to depressive news content in the month of February to drug use related news content in the month of March.

Using Drug Abuse-BERT model (Figure 7), states such as Texas, and Wisconsin shift from exposure of depressive news content in January to exposure of drug-related news content in February, while states such as California, and Oklahoma transition from exposure to depressive news content in February to drug-related news content in March. Further, we see the informativeness of news content sweeping from the east to the midwest, to parts of the south, and to some parts of the west from February to March.

Our results show that a fine-tuned BERT model cleanly separates the thematic categorical scores to a state. For instance, using DA-BERT for the month of March, the drug abuse score for the state of California is much higher than the score of depressiveness or informativeness for the same state. However, with the vanilla BERT model, the three scores computed for the various states and months are marginally different. Moreover, the results using DPR-BERT or DA-BERT capture the state-level ranking of mental disorders by Mental Health America [3] better than vanilla-BERT; for a few states, the fine-tuned BERT models identify more months to have media exposure to depression or drug abuse news content.

As indicated in Table 1, we report months showing predominant media exposure to either depressive or drug abuse news articles using the three variants of BERT model. We use 10 of the 13 states recognized as showing high prevalence of mental disorders according to a report by Mental Health America on overall mental disorder ranking. The 3 states not included in this table are Washington, Wyoming, and Idaho. We did not consider these 3 states as these states were not in our dataset cohort. For the Mental Health America (MHA) report, we make a practical assumption that each of the three months is either depressive or drug abusive for each state. Thus, our objective is to maximize the number of months with

[3] https://www.mhanational.org/issues/ranking-states

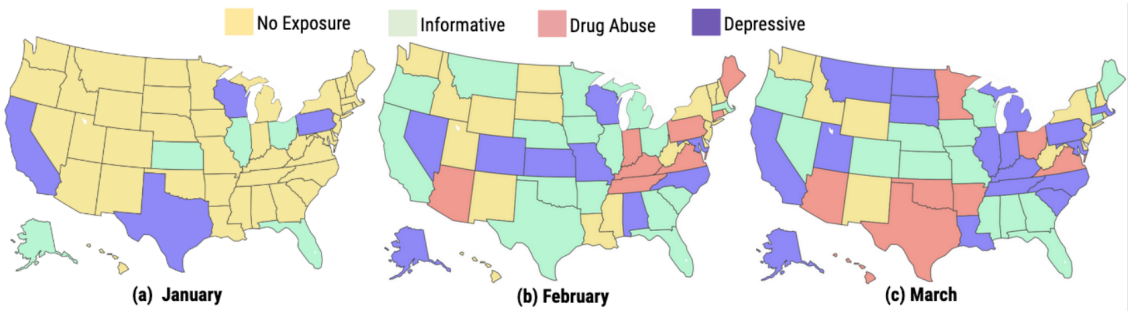


Figure 5: vanilla BERT modeling of Depressiveness, Drug Abuse, and Informativeness in US states.

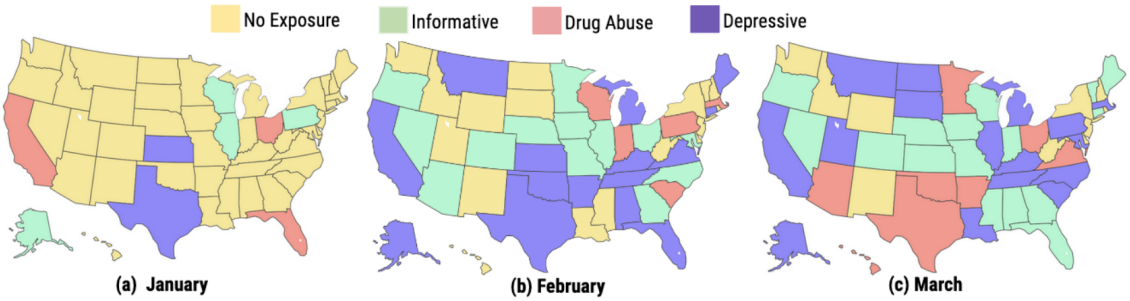


Figure 6: Depression-BERT (DPR-BERT) modeling of Depressiveness, Drug Abuse, and Informativeness in US states

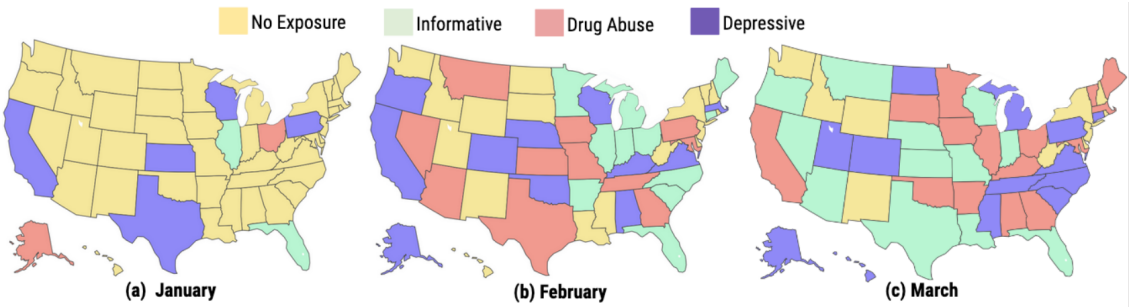


Figure 7: Drug Abuse BERT (DA-BERT) modeling of Depressiveness, Drug Abuse, and Informativeness in US states

exposure to depressive/drug abuse news content for each of the 10 states. We can see in Table 1 that fine-tuned BERT models help identify more months to having exposure to depressive or drug abuse news content than vanilla BERT does for the 10 states. For example, using DA-BERT, five states are identified to have at least two months showing exposure to depressive/drug abuse news content while DPR-BERT identifies six states to having been exposed to depressive/drug abuse news content for two months. On the other hand, vanilla-BERT identifies only two states with depressive/drug abuse news content for two months. To compare models with one another and against the report by Mental Health America (MHA), we compute a Jaccard Index between each pair of models and each model against the report from MHA. The equation below computes Jaccard similarity between the results of two models or a model's results with an MHA report.

$$J(m_1, m_2) = \sum_{i \in S}^{|S|} \frac{m_1^M \cap m_2^M}{m_1^M \cup m_2^M} \tag{2}$$

where,

$m_1, m_2 \in$ {vanilla-BERT, DPR-BERT, DA-BERT, MHA}

S - Set of States in the US (Table 1)

m_1^M, m_2^M: Number of depressive, drug abusive, or informative months for a state "i"

We report inter-model and model-to-MHA Jaccard similarity scores computed using equation (2) in Figure 8.

As shown in Figure 8, DA-BERT gives the best results against MHA report in Jaccard similarity (0.53), which means DA-BERT identifies over half of the state-to-month instances in MHA. On the other hand, vanilla-BERT has a Jaccard similarity of 0.37 with MHA, which can be interpreted as vanilla-BERT identifies a little over one-third of the state-to-month instances in MHA. The best Jaccard similarity is achieved between DPR-BERT and vanilla-BERT (0.7); thus, 70% of state-to-month mappings are shared between DPR-BERT and vanilla-BERT based on Jaccard index. It's interesting to see DA-BERT has the same Jaccard similarity with vanilla-BERT

MHA States with high DPR and DA	vanilla-BERT (Months with depression/drug abuse)	DA-BERT (Months with depression/drug abuse)	DPR-BERT (Months with depression/drug abuse)
Tennessee	Feb, Mar	Feb, Mar	Feb, Mar
Alabama	Feb	Feb, Mar	Feb
Oklahoma	Mar	Feb, Mar	Feb, Mar
Kansas	Feb	Jan, Feb	Jan, Feb
Montana	Mar	Feb	Feb, Mar
South Carolina	Mar	Mar	Feb, Mar
Alaska	Feb, Mar	Jan, Feb, Mar	Feb, Mar
Utah	Mar	Mar	Mar
Oregon	None	Feb	None
Nevada	Feb	Feb	None

Table 1: Evaluation of base and domain-specific BERT models for MHA states over the period of three months (January, February, and March). These three months showed high dynamicity in COVID-19 spread.

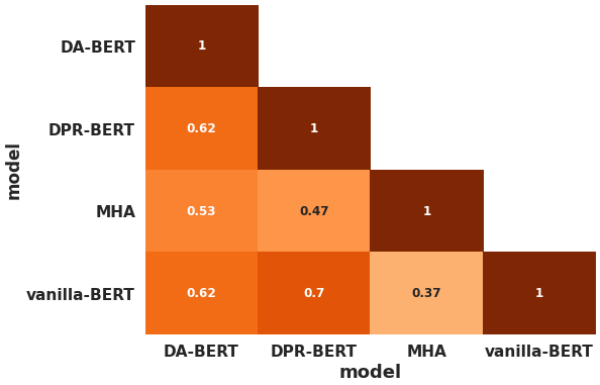


Figure 8: Inter-BERT model and BERT Model-to-MHA Jaccard Similarity Scores as a measure of closeness of model's prediction to an extensive survey on Mental Health America (MHA).

and DPR-BERT, subsuming the former and being subsumed by the latter in terms of depressive/drug abusive months.

5 CONCLUSION

In this paper, we model depressiveness, drug abusiveness, and informativeness of news articles to assess the dominant category characterizing each US state during each of the three months (Jan 2020 to Mar 2020). We demonstrate the power of transfer learning by fine-tuning an attention-based deep learning model on a different domain and use the domain-tuned model for gleaning the nature of media exposure. Specifically, we use background knowledge bases for measuring depressiveness, drug abusiveness, and informativeness of news articles. We found out DA-BERT identifies the most number of state-to-month instances as being exposed to depressive or drug abuse news content according to the report from Mental Health America. In the future, we plan to incorporate background knowledge bases in our attention-based transfer learning framework to further investigate knowledge-infused learning (Kursuncu et al., 2019).

REFERENCES

[1] Gennady Andrienko, Natalia Andrienko, Harald Bosch, Thomas Ertl, Georg Fuchs, Piotr Jankowski, and Dennis Thom. 2013. Thematic patterns in georeferenced tweets through space-time visual analytics. *Computing in Science & Engineering* 15, 3 (2013), 72–82.

[2] Delroy Cameron, Gary A Smith, Raminta Daniulaityte, Amit P Sheth, Drashti Dave, Lu Chen, Gaurish Anand, Robert Carlson, Kera Z Watkins, and Russel Falck. 2013. PREDOSE: a semantic web platform for drug abuse epidemiology using social media. *Journal of biomedical informatics* 46, 6 (2013), 985–997.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[4] Dana Rose Garfin, Roxane Cohen Silver, and E Alison Holman. 2020. The novel coronavirus (COVID-2019) outbreak: Amplification of public health consequences by media exposure. *Health psychology* (2020).

[5] Manas Gaur, Ugur Kursuncu, Amanuel Alambo, Amit Sheth, Raminta Daniulaityte, Krishnaprasad Thirunarayan, and Jyotishman Pathak. 2018. " Let Me Tell You About Your Mental Health!" Contextualized Classification of Reddit Posts to DSM-5 for Web-based Intervention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 753–762.

[6] George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim JP Hubbard, Richard JB Dobson, and Rina Dutta. 2017. Characterisation of mental health conditions in social media using Informed Deep Learning. *Scientific reports* 7 (2017), 45141.

[7] Benjamin Harbelot, Helbert Arenas, and Christophe Cruz. 2015. LC3: A spatio-temporal and semantic model for knowledge discovery from geospatial datasets. *Journal of Web Semantics* 35 (2015), 3–24.

[8] Emily A Holmes, Rory C O'Connor, V Hugh Perry, Irene Tracey, Simon Wessely, Louise Arseneault, Clive Ballard, Helen Christensen, Roxane Cohen Silver, Ian Everall, et al. 2020. Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. *The Lancet Psychiatry* (2020).

[9] Ugur Kursuncu, Manas Gaur, and Amit Sheth. 2019. Knowledge Infused Learning (K-IL): Towards Deep Incorporation of Knowledge in Deep Learning. *arXiv preprint arXiv:1912.00512* (2019).

[10] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 6, 2 (2015), 167–195.

[11] Meenakshi Nagarajan, Karthik Gomadam, Amit P Sheth, Ajith Ranabahu, Raghava Mutharaju, and Ashutosh Jadhav. 2009. Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. In *International Conference on Web Information Systems Engineering*. Springer, 539–553.

[12] Jianyin Qiu, Bin Shen, Min Zhao, Zhen Wang, Bin Xie, and Yifeng Xu. 2020. A nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic: implications and policy recommendations. *General psychiatry* 33, 2 (2020).

[13] Amit Sheth and Pavan Kapanipathi. 2016. Semantic filtering for social data. *IEEE Internet Computing* 20, 4 (2016), 74–78.

[14] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Npa: Neural news recommendation with personalized attention. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2576–2584.

[15] Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. 1191–1198.

Cost Aware Feature Elicitation

Srijita Das
The University of Texas at Dallas
Srijita.Das@utdallas.edu

Rishabh Iyer
The University of Texas at Dallas
Rishabh.Iyer@utdallas.edu

Sriraam Natarajan
The University of Texas at Dallas
Sriraam.Natarajan@utdallas.edu

ABSTRACT

Motivated by clinical tasks where acquiring certain features such as FMRI or blood tests can be expensive, we address the problem of test-time elicitation of features. We formulate the problem of cost-aware feature elicitation as an optimization problem with trade-off between performance and feature acquisition cost. Our experiments on three real-world medical tasks demonstrate the efficacy and effectiveness of our proposed approach in minimizing costs and maximizing performance.

CCS CONCEPTS

• **Supervised learning** → **Budgeted learning**; *Feature selection*; • **Applications** → Healthcare.

KEYWORDS

cost sensitive learning, supervised learning, classification

ACM Reference Format:
Srijita Das, Rishabh Iyer, and Sriraam Natarajan. 2020. Cost Aware Feature Elicitation. In *Proceedings of KDD Workshop on Knowledge-infused Mining and Learning (KiML'20).* , 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In supervised classification setting, every instance has a fixed feature vector and a discriminative function is learnt on such fixed-length feature vector and it's corresponding class variable. However, a lot of practical problems like healthcare, network domains, designing survey questionnaire [19, 20] etc has an associated feature acquisition cost. In such domains, there is a cost budget and getting all the features of an instance can be very costly. As a result, many cost sensitive classifier models [2, 8, 24] have been proposed in literature to incorporate the cost of acquisition into the model objective during training and prediction.

Our problem is motivated by such a cost-aware setting where the assumption is that prediction time features have an acquisition cost and adheres to a strict budget. Consider a patient visiting a doctor for some potential diagnosis of a disease. For such a patient, information like age, gender, ethnicity and other demographic features are easily available at zero cost. However, various lab tests that the patient needs to undergo incurs cost. So, a training model should be able to identify the most relevant (i.e. those which are most informative, yet least costly) lab tests that are required for each *specific* patient. The intuition of this work is that different patients, depending on their history, ethnicity, age and gender, may require different tests for reasonably accurate prediction. We build on the intuition that given certain observed features like one's demographic details, the most important features for a patient depends on the important features for similar patients. Based on this intuition, we find out similar data points in the observed feature space and identify the important feature subsets of these similar instances by employing a greedy information theoretic feature selector objective.

Our **contributions** in this work are as follows: (1) formalize the problem as a joint optimization problem of selecting the best feature subset for similar data points and optimizing the loss function using the important feature subsets. (2) account for acquisition cost in both the feature selector objective and classifier objective to balance the trade-off between acquisition cost and model performance. (3) empirically demonstrate the effectiveness of the proposed approach on three real-world medical data sets.

2 RELATED WORK

The related work on cost-sensitive feature selection and learning can be categorized into the following four broad approaches.

Tree based budgeted learning: Prediction time elicitation of features under a cost budget has been widely studied in literature. A lot of work has been done in tree based models [5, 16, 17, 26–28] by adding cost term to the tree objective function in either decision trees or ensemble methods like gradient boosted trees. All these methods aim to build an adaptive and complex decision tree boundary by considering trade-off between performance and test-time feature acquisition cost. While we are similar in motivation to these approaches, our methodology is different in the sense that we do not consider tree based models. Instead our approach aims to find local feature subsets using an information theoretic feature selector for different clusters of training instance build in a lower dimensional space.

Adaptive classification and dynamic feature discovery: Our work also draws inspiration from Nan al.'s work [15] where they learn a high performance costly model and approximate the model's performance adaptively by building a low cost model and gating function which decides which model to use for specific training instances. This adaptive switching between low and high cost model takes care of the trade-off between cost and performance. Our method is different from theirs because we do not maintain a high cost model which is costly to build and and difficult to decide. We refine the parameters of a single low cost model by incorporating a cost penalty in the feature selector and model objective. Our work is also along the direction of Nan et al.'s work [18] where they select varying feature subsets for test instance using neighbourhood information of the training data. While calculating the neighborhood information from training data is similar to building clusters in our approach, the training neighborhood for our method is on just the observed feature space. Moreover, we incorporate the neighbourhood information in the training algorithm whereas Nan et

In M. Gaur, A. Jaimes, F. Ozcan, S. Shah, A. Sheth, B. Srivastava, Proceedings of the Workshop on Knowledge-infused Mining and Learning (KDD-KiML 2020). San Diego, California, USA, August 24, 2020.
KiML'20, August 24, 2020, San Diego, California, USA,

https://doi.org/10.1145/nnnnnnn.nnnnnnn

al.'s work is a prediction time algorithm. Ma et al. [10] also address this problem of dynamic discovery of features based on generative modelling and Bayesian experimental design.

Feature elicitation using Reinforcement learning: There is another line of work along the sequential decision making literature [4, 9, 22] to model the test time elicitation of features by learning the optimal policy of test feature acquisition. Along this direction, our work aligns with the work of Shim et al. [25] where they jointly train a classifier and RL agent together. Their classifier objective function is similar to our method with a cost penalty, however they use a Deep RL agent to figure out the policy. We on the other hand use localised feature selector to find the important feature subsets for the underlying training clusters in the observed feature space.

Active Feature Acquisition: Our problem set-up is also inspired by work along active feature acquisition [13, 14, 19, 23, 29] where certain feature subsets are observed and rest are acquired at a cost. While all the above mentioned work follow this problem set up during training time and typically use active learning to seek informative instances at every iteration, we use this particular setting for test instances. Unlike their work, all the training instances in our work are fully observed and the assumption is that the feature acquisition cost has already being paid during training. Also, we address a supervised classification problem instead of an active learning set up. Our problem set up is similar to Kanani et al. [6] as they also have partial test instances, however their problem is that of instance acquisition where the acquired feature subset is fixed. Our method aims at discovering variable length feature subsets for various underlying clusters.

Our contributions: Although the problem of prediction time feature elicitation has been explored in literature from various directions and with various assumptions, we come up with an intuitive solution to this problem and formulate the problem in a **two step optimization framework**. We incorporate acquisition cost in both the feature selector and model objectives to balance the performance and cost trade-off. The problem set up is naturally applicable in real world health care and other domains where the knowledge of the observed features also needs to be accounted while selecting the elicitable features .

3 COST AWARE FEATURE ELICITATION

3.1 Problem setup

Given: A dataset $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ with each $x_i \in \mathbf{R}^d$ as the feature set. Each feature has an associated cost r_i.

Objective: Learn a discriminative model which is aware of the feature costs and can balance the trade-off between feature acquisition cost and model performance.

We make an additional assumption here that there is a subset of features which have 0 cost. These could be, for example, demographic information (e.g. age, gender, etc) in a medical domain which are easily available/less cumbersome to obtain as compared to other features. In other words, we can partition the feature set $\mathcal{F} = \mathcal{O} \cup \mathcal{E}$ where $\mathcal{O}$ are the zero cost observed features and $\mathcal{E}$ are the elicitable features which can be acquired at a cost. We also assume that the training data is completely available with all features (i.e. the cost for all the features has already been paid). The goal is to use these observed features to find similar instances in the training set and identify the important feature subsets for each of these clusters based on a feature selector objective function which balances the trade-off between choosing the important features and the cost at which these features are acquired.

3.2 Proposed solution

As a first step, we cluster the training instances based on just the observed zero cost feature set $\mathcal{O}$. The intuition is that instances with similar features will also have similar characteristics in terms of which elicitable features to order. For example, in a medical application, whether to request for a blood test or a ct-scan will depend on factors such as age, gender, ethnicity and whether patients with similar demographic features had requested these tests. Also, since the feature set $\mathcal{O}$, comes at zero cost, we assume that for unseen test instances, this feature set is observed.

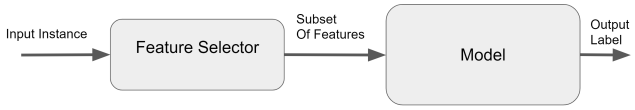


Figure 1: Optimization framework for the proposed problem

We propose a model which consists of a parameterized feature selector module $F(X, \mathbf{E}^{c_i}, \alpha)$ which takes in a set of input instances E^{c_i} belonging to the cluster c_i based on the feature set $\mathcal{O}$ and produces a subset X of most important features for the classification task. The feature selection model is based on an information- theoretic objective function and is augmented with the feature cost to account for the trade off between model performance and acquisition cost at test-time. The output feature subset from the feature selector module are used to update the parameters of the classifier. The optimization framework is shown in Figure 1

Information theoretic Feature selector model: The feature selector module selects the best subset of features for each cluster of training data based on an information theoretic objective score. Since at test time, we do not know the elicitable feature subset $\mathcal{E}$ (since the goal of feature selection is in the first place to find the truly necessary features for learning). Hence we propose to use the closest set of instances in the training data to the current instance. Since we assume that the training data has already been elicited, we have all the features observed in the training data. We compute this distance just based on the observed feature set $\mathcal{O}$. We cluster the training data based on the observed features into m clusters $c_1, c_2, \cdots c_m$. Next, we use the Minimum-Redundancy-Maximum Relevance (MRMR) feature Selection paradigm [1, 21]. We denote parameters $[\alpha^1_{c_i}, \alpha^2_{c_i}, \alpha^3_{c_i}, \alpha^4_{c_i}]$ as parameters of a particular cluster c_i. The feature selection module is a function of the parameters of

the cluster to which a set of instances belong and is defined as:

$$F(X, \mathbf{E}^{c_i}, \alpha_{c_i}) = \underbrace{\alpha^1_{c_i} \sum_{\mathcal{E}_p \in X} I(\mathcal{E}_p; Y)}_{\text{max. relevance}} - \underbrace{\sum_{\mathcal{E}_p \in X} \left(\alpha^2_{c_i} \sum_{\mathcal{E}_j \in X} I(\mathcal{E}_j; \mathcal{E}_p) - \alpha^3_{c_i} \sum_{\mathcal{E}_j \in X} I(\mathcal{E}_p; \mathcal{E}_j | Y) \right)}_{\text{min. redundancy}} - \underbrace{\alpha^4_{c_i} \sum_{\mathcal{E}_p \in X} c(\mathcal{E}_p)}_{\text{cost penalty}} \quad (1)$$

where $I(\mathcal{E}_p; Y)$ is the mutual information between the random variable $\mathcal{E}_p$ (feature) and Y (target). In the above equation, the feature subset X is grown greedily using a greedy optimization strategy maximizing the above objective function. In equation 1, $\mathcal{E}_p$ denotes a single feature from the elicitable set $\mathcal{E}$ that is considered for evaluation based on the subset X grown so far. The first term is the mutual information between each feature and the class variable Y. In a discriminative task, this value should be maximized. The second term is the pairwise mutual information between each feature to be evaluated and the features already added to the feature subset X. This value needs to be minimized for selecting informative features. The third term is called the conditional redundancy [1] and this term needs to be maximized. The last term adds the penalty for cost of every feature and ensures the right trade-off between cost, relevance and redundancy. For this work, we do not learn the parameters α_{c_i} for each cluster, instead fix these parameters to 1. We leave the learning of these parameters to future work.

In the problem setup, since the 0 cost feature subset is always present, we always consider the observed feature subset O in addition to the most important feature subset as returned by the Feature selector objective. We also account for the knowledge of the observed features while growing the informative feature subset through greedy optimization. Specifically, while calculating the pairwise mutual information between the features and the conditional redundancy term (second and third term of equation 1), we also evaluate the mutual information of the features with these observed features. It is to be noted that in cases where the observed features are not discriminative enough of the target, the feature selector module ensures that the elicitable features with **maximum relevance** to the target variable are picked.

Optimization Problem: The cost aware feature selector $F(X, \mathbf{E}^{c_i}, \alpha)$ for a given set of instance $\mathbf{E}^{c_i}$ belonging to a specific cluster c_i solves the following optimization problem:

$$X^i_\alpha = \text{argmax}_{X \subseteq \mathcal{E}} F(X, \mathbf{E}^{c_i}, \alpha) \quad (2)$$

For a given instance (x, y), we denote $L(x, y, X, \theta)$ as the loss function using a subset X of the features as obtained from the Feature selector optimization problem. The optimization problem for learning the parameters of a classifier can be posed as:

$$\min_\theta \sum_{i=1}^{n} L(x_i, y_i, X^i_\alpha, \theta) + \lambda_1 c(X^i_\alpha) + \lambda_2 ||\theta||^2 \quad (3)$$

where λ_1 and λ_2 are hyper-parameters. In the above equation, θ is the parameter of the model and can be updated by standard gradient based techniques. This loss function takes into account the important feature subset for each cluster and updates the parameter accordingly. The classifier objective also consists of a cost term denoted by $c(X^i_\alpha)$ to account for the cost of the selected feature subset. For hard budget on the elicited features, the cost component in the model objective can be considered. In case of a cost budget, this component can be ignored because the elicited feature subset adheres to a fixed cost and hence, this term is constant.

3.3 Algorithm

We present the algorithm for **Cost Aware Feature Elicitation** (CAFE) in Algorithm 1. CAFE takes as input set of training examples $\mathbf{E}$, the zero cost feature set O, the elicitable feature subset $\mathcal{E}$, a cost vector $M \in \mathbf{R}^d$ and a budget B. Each element in the training set $\mathbf{E}$ consists of a tuple (x, y) where $x \in \mathbf{R}^d$ is the feature vector and y is the label.

The training instances $\mathbf{E}$ are clustered based on just the observed feature set O using K-means clustering (`Cluster`). For every cluster c_i, the training instances belonging to the cluster is assigned to the set $\mathbf{E}^{c_i}$ and is passed to the Feature Selector module (lines 6-8). The `FeatureSelector` function takes $\mathbf{E}^{c_i}$, parameter α, the feature subsets O and $\mathcal{E}$, cost vector M and a predefined budget B as input and returns the most important feature subset $\mathbf{X}^{c_i}_\alpha$ corresponding to a cluster c_i. A greedy optimization technique is used to grow the feature subset X of every cluster based on the feature selector objective function defined in Equation 1. The `FeatureSelector` terminates once the budget B is exhausted or the mutual information score becomes negative. Once all the important feature subsets are obtained for all the $|C|$ clusters, the model objective function is optimized as mentioned in Equation 3 for all the training instances using the important feature subsets for the clusters to which the training instances belong (lines 12-18). All the remaining features are imputed by using either 0 or any other imputation model before training the model. The final training model $\mathtt{G}(\mathbf{E}_{O \cup X_\alpha}, \alpha, \theta)$ is an unified model used to make predictions for a test-instance consisting of just the observed feature subset O.

4 EMPIRICAL EVALUATION

We did experiments with 3 real world **medical data sets**. The intuition of CAFE makes more sense in medical domains, hence our choice of data sets. However, the idea can be applied to other domains ranging from logistics to resource allocation task. Table 2 jots down the various features of the data sets used in our experiments. Below are the details of the 3 real data sets, we use for our experiments.

1. **Parkinson's disease prediction:** The Parkinson's Progression Marker Initiative (PPMI) [12] is an observational study where the aim is to identify Parkinson's disease progression from various types of features. The PPMI data set consists of various features related to various motor functions and non-motor behavioral and psychological tests. We consider certain motor assessment features like rising from chair, gait, freezing of gait, posture and postural stability as observed features and rest all features as elicitable features which must be acquired at a cost.

Algorithm 1 Cost Aware Feature Elicitation

function CAFE($\mathbf{E}, O, \mathcal{E}, M, B$)
  $\mathbf{E} = \mathbf{E}_{O \cup \mathcal{E}}$ ▷ E consists of 0 cost features O and costly features $\mathcal{E}$
  C = Cluster($\mathbf{E}_O$) ▷ Clustering based on the observed features O
  $\mathbf{X} = \{\varnothing\}$ ▷ Stores best feature subsets of each cluster
  for $i = 1$ **to** $|C|$ **do** ▷ Repeat for every cluster
    $\mathbf{E}^{c_i}$ = GetClusterMember($\mathbf{E}, C, i$)
      ▷ get the data points belonging to each cluster c_i
    $\mathbf{X}_\alpha^{c_i}$ = FeatureSelector($\mathbf{E}^{c_i}, \alpha, O, \mathcal{E}, M, B$)
      ▷ Parameterized feature selector for each cluster
    $\mathbf{X} = \mathbf{X} \cup \{\mathbf{X}_\alpha^{c_i} \cup O\}$
  end for
  for $i = 1$ **to** $|C|$ **do** ▷ Repeat for every cluster
    $\mathbf{X}_\alpha^{c_i}$ = GetFeatureSubset($\mathbf{X}, i$)
      ▷ Get the feature subset for each cluster c_i
    for $j = 1$ **to** $|\mathbf{E}^{c_i}|$ **do** ▷ Repeat for every data point in cluster c_i
      Optimize $J(x_j, y_j, \mathbf{X}_\alpha^{c_i}, \theta, M)$
        ▷ Optimize the objective function in Equation 3
      Update θ ▷ Update the model parameter θ
    end for
  end for
  return G($\mathbf{E}_{O \cup X_\alpha}, \alpha, \theta$) ▷ G is the training model built on E
end function

2. **Alzheimer's disease prediction:** The Alzheimer's Disease NeuroIntiative (ADNI[1]) is a study that aims to test whether various clinical, FMRI and biomarkers can be used to predict the early onset of Alzheimer's disease. In this data set, we consider the demographics of the patients as observed and zero cost features and the FMRI image data and cognitive score data as unobserved and elicitable features.
3. **Rare disease prediction** This data set is created from survey questionnaires [11] and the task here is to predict whether a person has rare disease or not. The demographic features are observed while other sensitive questions in the survey regarding technology use, health and disease related meta information is considered to be elicitable.

Evaluation Methodology: All the data sets were partitioned into a 80:20 train-test split. Hyper parameters like the number of clusters on the observed features were picked by doing 5 fold cross validation on all the data sets. The optimal number of clusters picked were 6 for ADNI, 9 for Rare disease data set and 7 for the PPMI data set. For the results reported in Table 1, we considered a hard budget on the number of elicitable features and set it to half of the total number of features in the respective data set. We use K-means clustering as the underlying clustering algorithm. For all the reported results, we use an underlying Support Vector Machine [3] classifier with Radial basis kernel function. Since, all the data sets are highly imbalanced, hence we consider metrics like *recall*, *F1*, *AUC-ROC* and *precision* for our reported results. For the Feature selector module, we used the existing implementation of Li et al. [7]

[1]www.loni.ucla.edu/ADNI

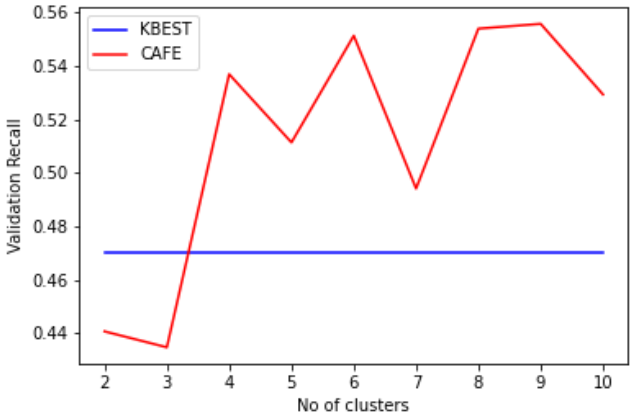


Figure 2: Recall Vs number of clusters for Rare disease for CAFE-I

and built upon it. We consider two variants of CAFE:(1) **CAFE** in which we replace the missing and unimportant features of every cluster with 0 and then update the classifier parameters (2) **CAFE-I** where we replace the missing and unimportant features by using an imputation model learnt from the already acquired feature values of other data points. A simple imputation model is used where we replace the missing features with *mode* for categorical features and *mean* for numeric features.

Baselines: We consider 3 baselines for evaluating CAFE and CAFE-I: (1) using the observed and zero cost features to update the training model denoted as OBS (2) using a random subset of fixed number of elicitable features and all the observed features to update the training model denoted as RANDOM. For this baseline, the results are averaged over 10 runs. (3) using the information theoretic feature selector score as defined in Equation 1 to select the 'k' best elicitable features on the entire data without any cluster consideration along with the observed features denoted as KBEST. We keep the value of 'k' to be the same as that used by CAFE. Although some of the existing methods could be potential baselines, none of these methods match the exact setting of our problem, hence we do not compare our method against them.

Results: We aim to answer the following questions:

Q1: How does **CAFE** and **CAFE-I** with hard budget on features compare against the standard baselines?
Q2: How does the cost-sensitive version of CAFE and CAFE-I fare against the cost-sensitive baseline KBEST?

The results reported in Table 1 suggests both CAFE and CAFE-I significantly outperform the other baselines in almost all the metrics for Rare disease and PPMI data set. For ADNI, CAFE and CAFE-I outperform the other baselines in clinically relevant recall metric while KBEST performs the best for the other metrics. The reason for this is that in ADNI, since, the elicitable features are image features and we discretize the image features to calculate the information gain for the feature selector module, the granular level feature information is lost because of this discretization and hence the drop in performance. For the experiments in Table 1, we keep the budget to be approximately half of the total number

Data set	Algorithm	Recall	F1	AUC-ROC	AUC-PR
Rare disease	OBS	**0.647**	0.488	0.642	0.347
	RANDOM	0.57 ± 0.064	0.549± 0.059	0.693 ± 0.042	0.421 ± 0.051
	KBEST	0.47	0.457	0.628	0.349
	CAFE	**0.647**	0.628	0.749	0.489
	CAFE-I	**0.647**	**0.647**	**0.759**	**0.512**
PPMI	OBS	0.765	0.685	0.741	0.563
	RANDOM	**0.857 ± 0.023**	0.809 ± 0.015	0.85 ± 0.013	0.712 ± 0.020
	KBEST	0.828	0.807	0.846	0.716
	CAFE	0.846	0.817	0.855	0.726
	CAFE-I	0.855	**0.829**	**0.865**	**0.743**
ADNI	OBS	0.5	0.44	0.553	0.365
	RANDOM	0.711 ± 0,043	0.697 ± 0.082	0.767 ± 0.064	0.592 ± 0.098
	KBEST	0.73	**0.745**	**0.806**	**0.646**
	CAFE	**0.807**	0.711	0.786	0.578
	CAFE-I	0.769	0.701	0.776	0.574

Table 1: Comparison of CAFE against other baseline methods on 3 real data sets

Dataset	# Pos	# Neg	# Observed	# Elicitable
PPMI	554	919	5	31
ADNI	94	287	6	69
Rare Disease	87	232	6	63

Table 2: Data set details of the 3 real data sets used.#Pos is number of positive example, #Neg is number of negative example. # Observed is number of observed features and # Elicitable is the maximum number of features that can be acquired.

of features for all the methods. On an average, CAFE-I performs better than CAFE across all the data sets because of the underlying imputation model which helps in better treatment of the missing values as against replacing all the features by 0. This answers **Q1** affirmatively.

In Figure 3, we compare the cost version of CAFE and CAFE-I against KBEST. Cost version takes into account the cost of individual features and accounts for them as penalty in the feature selector module. Hence, in this version of CAFE, a cost budget is used as opposed to hard budget on the number of elicitable features. We generate the cost vector by sampling each cost component uniformly from (0,1). For PPMI and Rare disease, we can see that cost sensitive CAFE performs consistently better than KBEST with increasing cost budget. In the PPMI data set, the greedy optimization of the feature selector objective on the entire data set lead to elicitation of just 1 feature, beyond that the information gain was negative, hence the performance of PPMI across various cross budget remains the same. CAFE on the other hand was able to select important feature subsets for various clusters based on the observed features related to gait and postures. For ADNI data set, CAFE performs better than KBEST only in recall. The reason for this is the same as mentioned above. This helps in answering **Q2** affirmatively.

Lastly, Figure 2 shows the effect of increasing cluster on the validation recall for the Rare disease data set. As can be seen, for smaller number of clusters, the recall is very low and increases to an optimum for 9 clusters. This helps us in understanding the fact that forming clusters based on observed important features helps CAFE in selecting different feature subsets for different clusters, thus helping the learning procedure.

5 CONCLUSION

In this paper, we pose the prediction time feature elicitation problem as an optimization problem by employing a cluster specific feature selector to choose the best feature subset and then optimizing the training loss. We show the effectiveness of our approach in real data sets where the problem set up is intuitive. Future work includes learning the parameters of the feature selector module and jointly optimizing the feature selector and model parameters for a more robust framework and adding more constraints to optimization.

ACKNOWLEDGEMENTS

SN & SD gratefully acknowledge the support of NSF grant IIS-1836565. Any opinions, findings and conclusion or recommendations are those of the authors and do not necessarily reflect the view of the US government.

REFERENCES

[1] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. 2012. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *JMLR* (2012).
[2] Xiaoyong Chai, Lin Deng, Qiang Yang, and Charles X Ling. 2004. Test-cost sensitive naive bayes classification. In *ICDM*.
[3] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* (1995).
[4] Gabriel Dulac-Arnold, Ludovic Denoyer, Philippe Preux, and Patrick Gallinari. 2011. Datum-wise classification: a sequential approach to sparsity. In *ECML PKDD*. 375–390.
[5] Tianshi Gao and Daphne Koller. 2011. Active classification based on value of classifier. In *NIPS*.
[6] P. Kanani and P. Melville. 2008. Prediction-time active feature-value acquisition for cost-effective customer targeting. *Workshop on Cost Sensitive Learning at NIPS* (2008).
[7] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. 2018. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)* (2018).
[8] Charles X Ling, Qiang Yang, Jianning Wang, and Shichao Zhang. 2004. Decision trees with minimal costs. In *ICML*.
[9] D. J. Lizotte, O. Madani, and R. Greiner. 2003. Budgeted learning of Naive-Bayes classifiers *(UAI)*. 378–385.
[10] Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez-Lobato, Sebastian Nowozin, and Cheng Zhang. 2019. EDDI: Efficient Dynamic Discovery of High-Value Information with Partial VAE. In *ICML*.
[11] H. MacLeod, S. Yang, et al. 2016. Identifying rare diseases from behavioural data: a machine learning approach *(CHASE)*. 130–139.
[12] K. Marek, D. Jennings, et al. 2011. The Parkinson Progression Marker Initiative (PPMI). *Prog Neurobiol* 95, 4 (2011), 629–635.

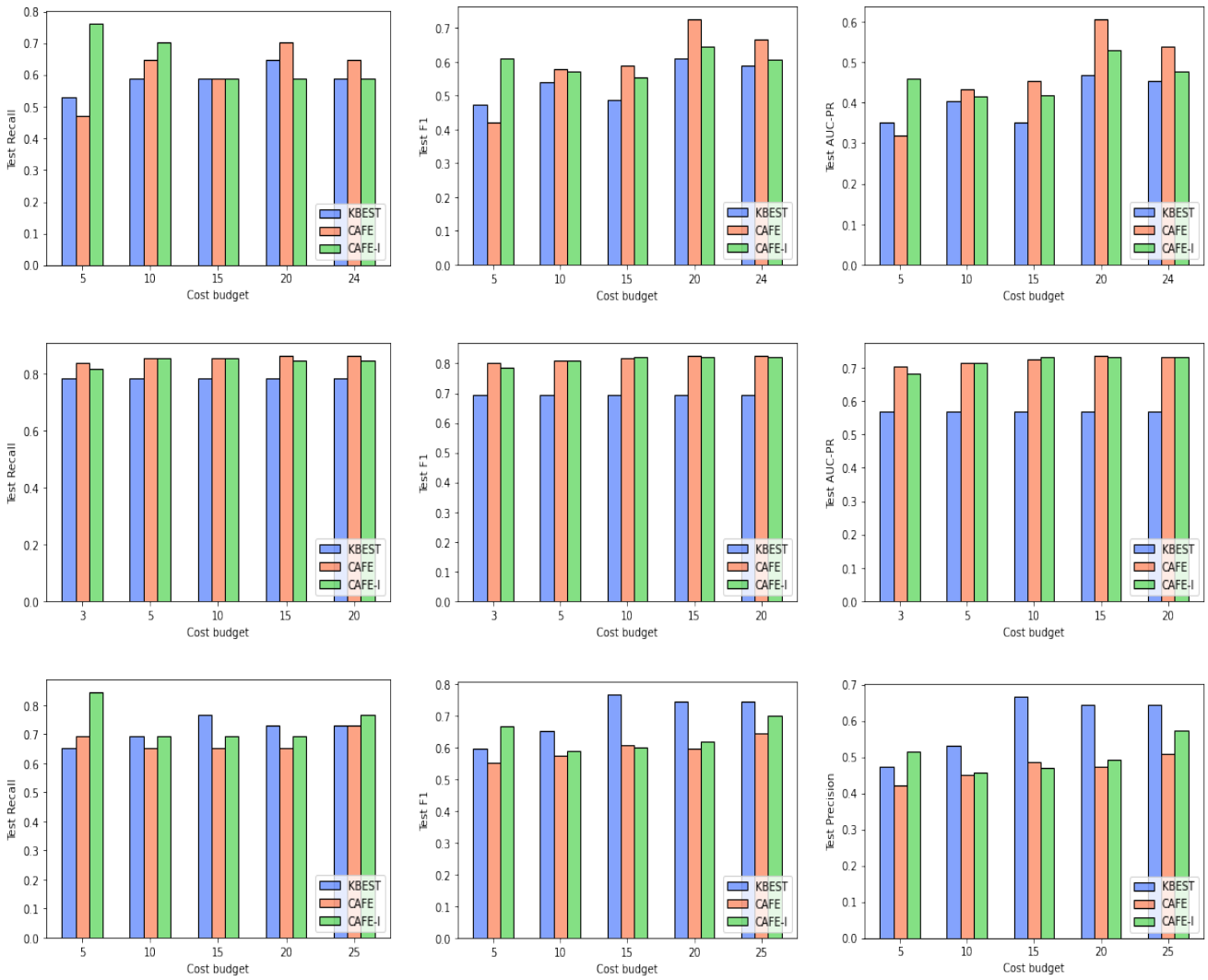


Figure 3: Recall (left), F1 (middle), AUC-PR (right) for (from top to bottom) Rare Disaese, PPMI, and ADNI. The x-axis refers to the cost budget used which leads to the elicitation of different number of features.

[13] P. Melville, M. Saar-Tsechansky, et al. 2004. Active feature-value acquisition for classifier induction *(ICDM)*. 483–486.

[14] P. Melville, M. Saar-Tsechansky, et al. 2005. An expected utility approach to active feature-value acquisition *(ICDM)*. 745–748.

[15] Feng Nan and Venkatesh Saligrama. 2017. Adaptive classification for prediction under a budget. In *NIPS*.

[16] Feng Nan, Joseph Wang, and Venkatesh Saligrama. 2015. Feature-budgeted random forest. In *ICML*.

[17] Feng Nan, Joseph Wang, and Venkatesh Saligrama. 2016. Pruning random forests for prediction on a budget. In *NIPS*.

[18] Feng Nan, Joseph Wang, Kirill Trapeznikov, and Venkatesh Saligrama. 2014. Fast margin-based cost-sensitive classification. In *ICASSP*.

[19] Sriraam Natarajan, Srijita Das, Nandini Ramanan, Gautam Kunapuli, and Predrag Radivojac. 2018. On Whom Should I Perform this Lab Test Next? An Active Feature Elicitation Approach.. In *IJCAI*.

[20] S. Natarajan, A. Prabhakar, et al. 2017. Boosting for postpartum depression prediction *(CHASE)*. 232–240.

[21] Hanchuan Peng, Fuhui Long, and Chris Ding. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence* 27, 8 (2005), 1226–1238.

[22] Thomas Rückstieß, Christian Osendorfer, and Patrick van der Smagt. 2011. Sequential feature selection for classification. In *Australasian Joint Conference on Artificial Intelligence*. Springer, 132–141.

[23] M. Saar-Tsechansky, P. Melville, and F. Provost. 2009. Active feature-value acquisition. *Manag Sci* 55, 4 (2009).

[24] Victor S Sheng and Charles X Ling. 2006. Feature value acquisition in testing: a sequential batch test algorithm. In *ICML*.

[25] Hajin Shim, Sung Ju Hwang, and Eunho Yang. 2018. Joint active feature acquisition and classification with variable-size set encoding. In *NIPS*.

[26] Joseph Wang, Kirill Trapeznikov, and Venkatesh Saligrama. 2015. Efficient learning by directed acyclic graph for resource constrained prediction. In *NIPS*.

[27] Zhixiang Xu, Matt Kusner, Kilian Weinberger, and Minmin Chen. 2013. Cost-sensitive tree of classifiers. In *ICML*.

[28] Zhixiang Xu, Kilian Q Weinberger, and Olivier Chapelle. 2012. The greedy miser: learning under test-time budgets. In *ICML*.

[29] Z. Zheng and B. Padmanabhan. 2002. On active learning for data acquisition *(ICDM)*. 562–569.

A New Delay Differential Equation Model for COVID-19

Retarded logistic equation

B Shayak†
Mechanical and Aerospace Engg
Cornell University
Ithaca, New York State, USA
sb2344@cornell.edu

Mohit M Sharma
Population and Health Sciences
Weill Cornell Medicine
New York City, USA
mos4004@med.cornell.edu

Manas Gaur
AI Institute
University of South Carolina
USA
mgaur@email.sc.edu

ABSTRACT

In this work we give a delay differential equation, the retarded logistic equation, as a mathematical model for the global transmission of COVID-19. This model accounts for asymptomatic carriers, pre-symptomatic or latent transmission as well as contact tracing and quarantine of suspected cases. We find that the equation admits varied classes of solutions including self-burnout, progression to herd immunity and multiple states in between. We use the term "partial herd immunity" to refer to these states, where the disease ends at an infection fraction which is not negligible but is significantly lower than the conventional herd immunity threshold. We believe that the spread of COVID-19 in every localized area can be explained by one of our solution classes.

CCS CONCEPTS

• Applied computing – mathematics and statistics

KEYWORDS

Retarded logistic equation, Asymptomatic carriers, Latent transmission, Contact tracing, Reproduction number calculation, Partial herd immunity

1 Introduction

Three kinds of models to study COVID-19 are currently in vogue – lumped parameter or compartmental models (ordinary differential equation), agent-based models and stochastic differential equation models. The first option affords maximum conceptual clarity at the expense of some simplifying assumptions (homogeneous mixing etc). The second option affords maximum potential versatility at the cost of huge computational complexity and variability in the network structure. The third option combines features of the previous two – whether the features being synergized are the positive or the negative ones depends to a large extent on the modeler.

In this work we use delay differential equations (DDE) to propose a simple, single-variable, lumped parameter model for the spread of Coronavirus. Jahedi and Yorke [1] make a strong case for simpler models relative to complex and elaborate ones. In the Literature, DDE has been used for modeling COVID-19, for example in Refs. [2]–[4]. These authors however ignore features such as contact tracing, asymptomatic carriers and latent transmission; our results too have a richer structure.

2 Derivation of the model

We measure time t in days and use as our basic variable $y(t)$ which is the cumulative number of corona cases, including active cases, recovered cases and deaths, in the region of interest. The following "word-equation" summarizes the approach :

$$\begin{pmatrix}\text{Rate of emergence}\\ \text{of new cases}\end{pmatrix} = \begin{pmatrix}\text{Interaction rate of}\\ \text{each existing case}\end{pmatrix} \times \begin{pmatrix}\text{Probability of}\\ \text{transmission}\end{pmatrix} \times \begin{pmatrix}\text{Number of}\\ \text{existing cases}\end{pmatrix} \tag{0}$$

The left hand side (LHS) here is just $\mathrm{d}y/\mathrm{d}t$ whereas the right hand side (RHS) needs a detailed derivation.

Equation (0) assumes that the disease is transmitted from infected to susceptible people via interaction, and not via airborne transmission. Due to asymptomatic and pre-symptomatic carriers, there are always cases moving about in society who are oblivious to their infectivity. Each such case interacts with other people at a different rate. For example, a working-from-home professor might venture outside once every three days and interact with one person on each trip while a grocer might go to work and interact with 10 customers every day. The professor has an interaction rate of 1/3 persons/day while the grocer has interaction rate of 10

†Presenting author, Corresponding author. ORCID : 0000-0003-2502-2268

In M. Gaur, A. Jaimes, F. Ozcan, S. Shah, A. Sheth, B. Srivastava, Proceedings of the Workshop on Knowledge-infused Mining and Learning (KDD-KiML 2020). San Diego, California, USA, August 24, 2020.

KiML'20, San Diego, California, USA

persons/day. For a compartmental model, one must average over the professor, the grocer and all the other un-quarantined cases to generate an effective per-case interaction rate q_0.

Every interaction of course does not result in a transmission – there is a probability strictly less than unity that the virus jumps from the infected person to the person whom s/he is interacting with. This probability has two components. The first component is that the healthy person must be susceptible to begin with. While we ignore intrinsic insusceptibles, there will be people who have recovered from the disease and are therefore not susceptible again. In this Article, we assume that one bout of infection brings permanent immunity. The assumption is valid so long as the immunity period exceeds the total epidemic duration. Till date, there is little credible evidence for re-infection [5]–[7]; contrarily, a very recent and thorough study [8] based on monitoring of huge patient cohort has found significant evidence of long-lasting and effective antibodies. If N be the initial number of susceptible people (recall that y is the case count), then the probability that a random person is a recovered case is approximately y/N and the probability that s/he is susceptible is (approximately) $1-y/N$. This expression is approximate because the true number of recovered cases at any time is less than y; the error however is small since the recovery period is much shorter than the overall course of the epidemic. Note that $1-y/N$ is a logistic term, and a herd immunity effect.

Given susceptibility, the next probability is that the virus actually does jump from the un-quarantined case to the susceptible person. This probability depends on the level of precaution such as face covering or mask, handwashing and disinfection being adopted by the case as well as the susceptible person. For a compartmental model, the probability must be averaged over all the un-quarantined cases. If this average probability is P_0, then $q_0(1-y/N)P_0$ gives the per-case spreading rate. Since q_0 and P_0 are both dependent on public health measures, and are both difficult to measure independently, we can club those two together into a single parameter which we call m_0.

So far we have accounted for the rate at which each cases spreads the disease; now we have to count the number of cases out of quarantine. Let us start with an asymptomatic carrier, who remains in open society throughout. S/he typically transmits the disease for 7 days, which is called the infection period. Then, new healthy people can be only be infected by those asymptomatic cases who have fallen sick within the last 7 days, and not those who have fallen sick earlier. The number of such people is the number of asymptomatic sick people today minus the number of those 7 days earlier. Mathematically, let μ_1 (between 0 and 1) denote the fraction of asymptomatic carriers and τ_1 the asymptomatic infection period. Then, the number of asymptomatic transmitters today is $\mu_1(y(t)-y(t-\tau_1))$. Here we can see the emergence of the delay term.

The remaining fraction $1-\mu_1$ of cases are symptomatic. Let τ_2 be the latency period during which these cases remain transmissible prior to displaying symptoms. It is assumed that they isolate themselves thereafter. Assumption is also made that the incubation period is equal to the latency period. Finally, the contact tracing drive conducted by public health department is taken into account. Assumption is made that this drive is instantaneous and proceeds in forward direction starting from freshly arriving symptomatic cases. The contact trace captures patients who were exposed to the new case τ_2 days ago, as well as patients who were exposed immediately before the new case manifested symptoms. The average duration for which these secondary patients have remained at large is $\tau_2/2$, be they symptomatic or asymptomatic. The assumption of instantaneous contact tracing, which decreases the average time that contact-traced cases spend out of quarantine, opposes the error arising from the assumption of zero non-transmissible incubation period, which increases the average time for which the contact-traced cases transmit before quarantine. These two effects are assumed here to cancel. Let μ_3 (between 0 and 1) denote the fraction of all cases who escape from contact tracing drives – the complementary fraction $1-\mu_3$ get caught. Thus, we have three classes of un-quarantined cases : (*a*) $1-\mu_3$ are contact-traced cases who remain in society for a time $\tau_2/2$, (*b*) μ_3 $(1-\mu_1)$ are untraced symptomatic cases who go into isolation only after time τ_2, and (c) $\mu_3\mu_1$ are undetected asymptomatic cases who transmit for the entire infection period τ_1. Arguments similar to those of the previous paragraph yield the total number of un-quarantined cases as

$$n = (1-\mu_3)(y - y(t-\tau_2/2)) + ((1-\mu_1)\mu_3)(y - y(t-\tau_2)) + \mu_1\mu_3(y - y(t-\tau_1)) \quad . (1)$$

The preceding arguments now yield the mathematical form of (0) as

$$\frac{dy}{dt} = m_0\left[1-\frac{y}{N}\right]\left[y(t) - (1-\mu_3)y(t-\tau_2/2) - (1-\mu_1)\mu_3 y(t-\tau_2) - \mu_1\mu_3 y(t-\tau_1)\right] \quad (2)$$

which is the retarded logistic equation.

3 Solutions of the model

Due to complexity of the equation (2), analytical solution using perturbation theory etc has not been attempted in this case. Instead we have used numerical integration to obtain the solutions of (2). Before giving the solutions however, we present the calculation of the reproduction number R. To find R at any state of evolution of the disease, we first treat y in the logistic term to be constant, and then carry out the steps described in Ref. [9]. This yields the expression

$$R = m_0\left(1-\frac{y}{N}\right)\left(\frac{1+\mu_3-2\mu_1\mu_3}{2}\tau_2 + \mu_1\mu_3\tau_1\right) \quad . (3)$$

The ease of calculating R with respect to the ordinary differential equation based models [10] is noteworthy.

Solution classes of logistic DDE (2) are now demonstrated. The numerical integration routine used is second order Runge Kutta with a time step of 1/1000 day. As the testbed for the simulations, we consider a Notional City having N=300000, μ_1=0.8, (maximum

value as per our knowledge [11]–[13]), τ_1=7 days and τ_2=3 days [14]. The initial condition needs to be a function having the length of the maximum delay involved in the problem, which is seven days; we take this function to be zero cases to start with and constant increase of 100 cases/ day for a week.

Notional City A has m_0=0.23 and μ_3=1/2, which describes a hard lockdown [15] accompanied by good contact tracing. R_0 (i.e. (3) evaluated at y=0) is 0.886. The epidemic ends with a negligible fraction of infected people, as shown below. This and the next five plots are three-way – each plot shows y as blue line, its derivative $\dot{y}$ as green line and the weekly increments in cases, or epidemiological curve, as a grey bar chart. These last have been reduced by a factor of 7 to ensure clarity of presentation. We report the rates on the left hand side y-axis and the cumulative cases on the right hand side y-axis.

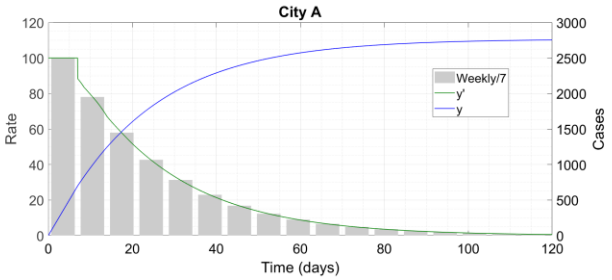


Figure 1 : City A extinguishes the epidemic in time.

This is exactly what has happened in New Zealand – that il fortunatissimo per verita has indeed quashed the epidemic completely with the final case count being a negligible fraction of its total (tiny and sparsely distributed) population.

The parameter values for Notional City B are the same as those for A except that μ_3=0.75; a greater fraction of cases escape the contact tracing drive. R_0 is 1.16, and R becomes 1 at y=40500 cases.

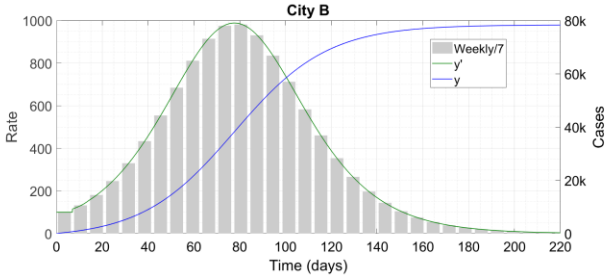


Figure 2 : City B grows at first before reaching burnout. The symbol 'k' denotes thousand.

The outbreak enters exponential regime right after being released. As y increases, R gradually reduces so the growth slows down until it peaks when the case count is about 39,000 [compare with the value of 40,500 when R=1 as per (3)]. Thereafter, the disease progresses to extinction in time. The overall progression is very long but one hopes that the relatively small size of the peak can prevent overstressing of medical care facilities and thus avoid unnecessary deaths. Delhi and Mumbai in India and Los Angeles in USA are in all probability cities of this type since the disease there spiraled out of control despite hard lockdowns being imposed at an early stage.

City B also enables us to explain partial herd immunity. Even though the initial conditions were unfavourable for containment of the epidemic, herd immunity started activating as the disease proliferated. A stable zone (R<1) was entered when only 13.5 percent of the total susceptible population was infected, and a similar percentage again got infected before the epidemic ended. Thus, herd immunity worked in synergy with non-pharmaceutical interventions to stop the epidemic at only 26 percent infection level, which is significantly less than the conventional 70-90 percent threshold [16]. This is what we call partial herd immunity. Our findings are in agreement with and act as an explanation for what has been obtained by Britton et. al. [17] and Peterson et. al. [18].

We now consider Notional City C which differs from City B in that m_0=0.5; lockdown is replaced by a much more permissive state. R_0 is above 2.5; 1,80,000 infections are required to bring it below unity.

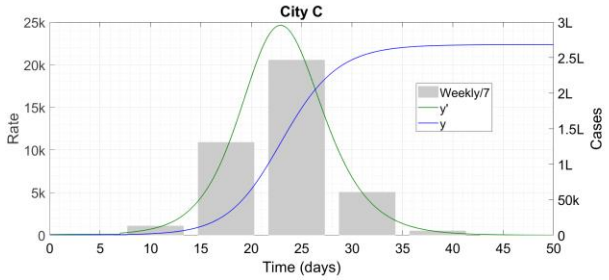


Figure 3 : City C goes to herd immunity – total not partial. The symbol 'k' denotes thousand and 'L' hundred thousand.

Need one mention that this is a public health disaster. Notional City D combines features of B and C. This city begins with m_0=0.5 like City C but reduces to m_0=0.23 like City B when the case count reaches 40,000 (the R=1 threshold for B's parameters).

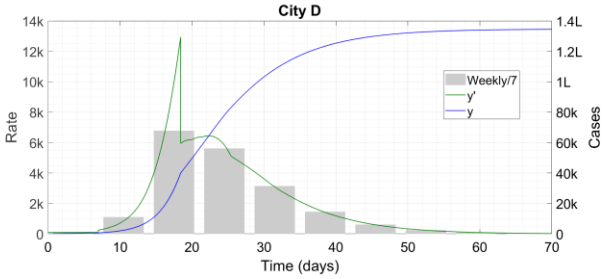


Figure 4 : As the input, so the output – D's response combines features of B and C. The symbol 'k' denotes thousand and 'L' hundred thousand.

We can see a case count as well as a total duration intermediate to B and C; the epidemic is over in 70 days but the peak rate of 12,920 cases/day is still very high and likely to load hospital facilities beyond their carrying capacity.

The Cities E and F demonstrate the issues faced in reopening. In both these cities, the parameters and case trajectory are

identical to those of City A for the first 80 days. Then, E and F reopen on the 80th day by increasing m_0 from 0.23 to 0.5, and simultaneously decreasing μ_3 i.e. deploying a more effective contact tracing program which had been built up during the lockdown. The post-reopening μ_3's for E and F are 0.1 and 0.2 respectively.

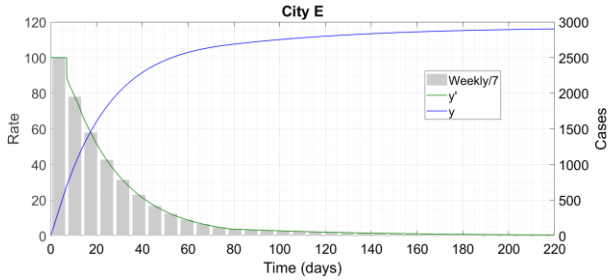


Figure 5 : City E, like City A, is a success story.

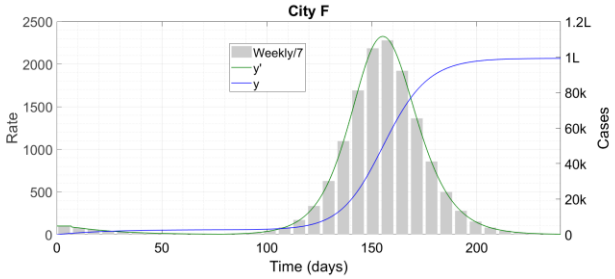


Figure 6 : Unlike City E, F is a failure story. The symbol 'k' denotes thousand and 'L' hundred thousand.

The difference between Cities E and F is dramatic. Mathematically, R remained less than unity throughout in E; its value after reopening was 0.985. We can see that the case rate decreases monotonically all the time. In F, the post-reopening R became 1.22 and sent the trajectory haywire. In practice however, the incipient increase in case rate after the 80th day acts as an advance warning of what has happened – the reopening steps should be reversed if it is at all possible to do so while satisfying economic and other external constraints.

Conclusion

In this Article we have presented a new mathematical model for COVID-19 which is simple and elegant in structure but can generate a variety of realistic solution classes. We hope that our work may be of use to mathematicians and data scientists who are trying to understand the spread of the disease in a quantitative manner. The public health implications of these results are being reserved for another study.

REFERENCES

[1] S. Jahedi and J. A. Yorke, "When the best pandemic models are the simplest .," *medRxiv*, pp. 1–22, 2020, doi: https://doi.org/10.1101/2020.06.23.20132522.

[2] L. Dell'Anna, "Solvable delay model for epidemic spreading: the case of Covid-19 in Italy," 2020, [Online]. Available: http://arxiv.org/abs/2003.13571.

[3] A. K. Gupta, N. Sharma, and A. K. Verma, "Spatial Network based model forecasting transmission and control of COVID-19," *medRxiv*, p. 2020.05.06.20092858, 2020, doi: 10.1101/2020.05.06.20092858.

[4] J. Mendenez, "Elementary time-delay dynamics of COVID-19 disease," *Medrxiv*, pp. 1–4, 2020, doi: https://doi.org/10.1101/2020.03.27.20045328.

[5] D. C. Ackerly, "Getting COVID-19 twice." VOX, [Online]. Available: https://www.vox.com/2020/7/12/21321653/getting-covid-19-twice-reinfection-antibody-herd-immunity.

[6] S. McCamon, "13 USS Roosevelt Sailors Test Positive For COVID-19, Again."

[7] Y. Saplakoglu, "coronavirus-reinfections-were-false-positives." [Online]. Available: https://www.livescience.com/coronavirus-reinfections-were-false-positives.html.

[8] A. Wajnberg *et al.*, "SARS-CoV-2 infection induces robust, neutralizing antibody responses that are stable for at least three months," *medRxiv*, 2020, doi: https://doi.org/10.1101/2020.07.14.20151126.

[9] B. Shayak and R. H. Rand, "Self-burnout - A New Path to the End of COVID-19," *medRxiv*, pp. 1–14, 2020, doi: https://doi.org/10.1101/2020.04.17.20069443.

[10] O. Diekmann, J. A. P. Heesterbeek, and M. G. Roberts, "The construction of next-generation matrices for compartmental epidemic models," *J. R. Soc. Interface*, vol. 7, no. 47, pp. 873–885, 2010, doi: 10.1098/rsif.2009.0386.

[11] "71 percent of patients in Maharashtra are asymptomatic." Mumbai Mirror, [Online]. Available: https://mumbaimirror.indiatimes.com/coronavirus/news/covid-19-71-of-patients-in-maharashtra-are-asymptomatic-mumbai-cases-at-16579/articleshow/75754328.cms.

[12] "Taking over hospital beds, conducting survey." New Indain Express, [Online]. Available: https://www.newindianexpress.com/nation/2020/may/30/taking-over-hospital-beds-conducting-survey-uddhav-government-goes-after-covid-19-as-state-tally-c-2149989.html.

[13] "Delhi CM says COVID-19 deaths very less." Times of India, [Online]. Available: https://timesofindia.indiatimes.com/city/delhi/delhi-cm-says-covid-19-deaths-very-less-but-75pc-cases-asymptomatic-or-showing-mild-symptoms/articleshow/75658636.cms.

[14] M. L. Childs *et al.*, "The impact of long-term non-pharmaceutical interventions on COVID-19 epidemic dynamics and control," *medRxiv*, vol. 22, p. 2020.05.03.20089078, 2020, doi: 10.1101/2020.05.03.20089078.

[15] B. Shayak and M. M. Sharma, "Retarded Logistic Equation as a Universal Dynamic Model for the Spread of COVID-19," *medRxiv*, pp. 1–27, 2020, doi: 10.1101/2020.06.09.20126573.

[16] G. A. D'Souza and D. Dowdy, "What is herd immunity and how we can achieve it with COVID-19?" [Online]. Available: https://www.jhsph.edu/covid-19/articles/achieving-herd-immunity-with-covid19.html.

[17] T. Britton, F. Ball, and P. Trapman, "The disease-induced herd immunity level for Covid-19 is substantially lower than the classical herd immunity level," pp. 1–15, 2020, [Online]. Available: http://arxiv.org/abs/2005.03085.

[18] A. A. Peterson, C. F. Goldsmith, C. Rose, A. J. Medford, and T. Vegge, "Should the rate term in the basic epidemiology models be second-order?," 2020, [Online]. Available: http://arxiv.org/abs/2005.04704.

Public Health Implications of a delay differential equation model for COVID 19

Mohit M Sharma
Population and Health Sciences
Weill Cornell Medicine
New York City, USA

mos4004@med.cornell.edu

B Shayak
Sibley School of Mechanical and Aerospace Engineering
Cornell University
Ithaca, New York State, USA

sb2344@cornell.edu

ABSTRACT

This paper describes the strategies derived from a novel delay differential equation model[1], signifying a practical extension of our recent work. COVID -19 is an extremely ferocious and an unpredictable pandemic which poses unique challenges for public health authorities, on account of which "case races" among various countries and states do not serve any purpose and present delusive appearances while ignoring significant determinants. We aim to propose comprehensive planning guidelines as a direct implication of our model. Our first consideration is reopening, followed by effective contact tracing and ensuring public compliance. We then discuss the implications of the mathematical results on people's behavior and eventually provide conclusive points aimed at strengthening the arsenal of resources that are helpful in framing public health policies. The knowledge about pandemic and its association with public health interventions is documented in the various literature-based sources. In this study, we explore those resources to explain the findings inferred from delay differential equation model of covid-19.

KEYWORDS

Delay differential equation, Contact tracing, Socio-behavioral theories, Lockdown, Reopening

1 INTRODUCTION

The national (USA) and global spread of Coronavirus Disease 2019 (COVID-19), following its origins in Wuhan, China in at least December 2019 and possibly earlier still [2] has been alarmingly rapid and deadly. From the 25 individual national forecasts received by CDC, predicts that there is possibility of the total reported COVID -19 deaths is between 160,000 and 175,000 by August 15th, 2020 [3]. Some features however, both nationally and globally, have proved counterintuitive. For example, a 76-day lockdown resulted in the outbreak's containment in Wuhan. A similar measure has produced similar results in New Zealand. However, lockdown appeared only marginally effective in New York State, USA where the case and death counts decreased only after reaching horrifying peak levels [4]. It was contended that the stay at home order in New York came too late. This apparent delay was not present in California, USA. The case counts there went up all the same, and the rate is high even today. We would like to mention that such spatiotemporal anomalies are present not just in the US but also in other countries such as Canada, Russia and India [5] which witnessed high case growth despite being in lockdown. In order to better understand the epidemiology of the transmission of COVID-19, we have constructed a delay differential equation model. Here we present its practical implications which tries to encapsulate a myriad of factors associated with the current scenario.

2 MATHEMATICAL MODELING TO UNDERSTAND THE EPIDEMIOLOGY

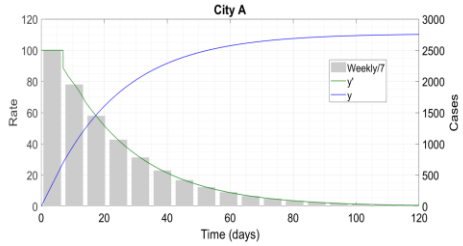
Since many decades, mathematical modelling has been used as an integral tool in recognizing the trend of disease progression during pandemics. For example, using a simple model explaining the transmission dynamics of the infectious disease between the susceptible, infected and recovered population (SIR Epidemic Models) Kermack and McKendrick proposed and later established a principle – the level of susceptibility in the population should be adequately high in order for that epidemic to unfold in that population. Such mathematical models can give impressionable insights in explaining the epidemiological status of the population, predict or calculate the transmissibility of the pathogen and the potential impact of public health preventive

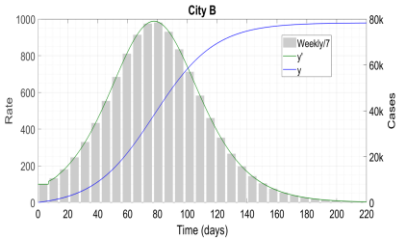
In M. Gaur, A. Jaimes, F. Ozcan, S. Shah, A. Sheth, B. Srivastava, Proceedings of the Workshop on Knowledge-infused Mining and Learning (KDD-KiML 2020). San Diego, California, USA, August 24, 2020.
KiML'20, San Diego, California, USA

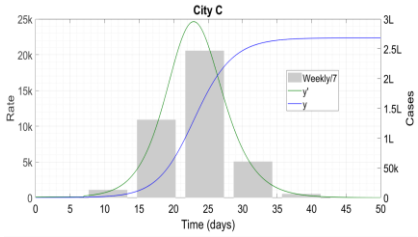
practices [6]. However, a significant body of evidence suggests that decisions should be made regarding the parameters to be included, being contingent on the impact of the precision of predictions. Several policy questions about the containment of this outbreak have been considered in our recently proposed simple non-linear model [1]This paper delves into the practical solutions that can be devised utilizing the directions of our models' outcome.

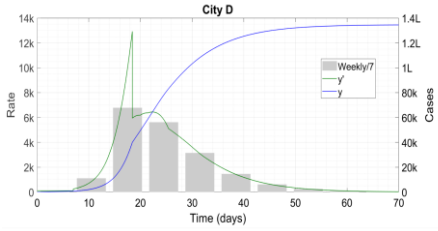
In generating interpretable results gathered from epidemiological models, we have used the examples of six types of cities [1]:

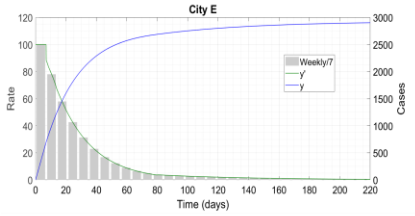
1) City A – Moderately effective contact tracing in a hard lockdown. This city has R (reproductive number) <1 and drives epidemic to extinction in time.
2) City B – Less effective contact tracing in a hard lockdown. It starts off R >1, but reached R =1 at 15% infection level. The epidemic ends at 30% infection rate and takes a very long time to get there.
3) City C – Less effective contact tracing (Like City B) with milder restriction on mobility. It proceeds rapidly to herd immunity.
4) City D – Combination of City B and City C. Starts with mild restriction on mobility and progresses towards restriction. The duration of the epidemic as well as of the final case count is between CITIES B and C.
5) City E - Starts off like City A, it reopens with very effective contact tracing and drive the epidemic to extinction in time.
6) City F – Starts off like CITY A, it reopens with less effective contact tracing and suffers a second wave.

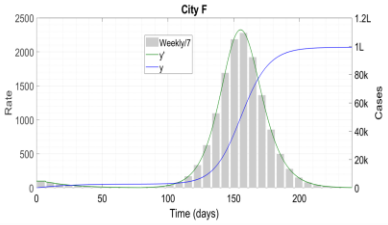












Pragmatic implications of our work are as follows:

3 REOPENING CONSIDERATIONS, ROLE OF TESTING

The unemployment situation generated as a result of lockdowns is currently forcing countries and states to partially reopen their economies even though many of them have not yet got the virus under control. The reopening is easiest in City A regions where cases have slowed down to a trickle. With every new case being detected, swift isolation of all potential secondary, tertiary and maybe even quaternary cases, both forward and backward, should prove possible while the rest of the economy functions in a relatively uninhibited way. Even one mass transmission event can restart an exponential growth regime and force a rollback to a fully locked down state. Reopening beyond a skeletal level is impossible in City B regions which are still in the ascending phase. The ascent implies that contact tracing is already inadequate, and on top of that if mobility increases then the region might turn into City C, overstress healthcare systems, and become a massacre. An ascending B-City has little option other than to contact trace as hard as possible and wait for partial herd immunity to kick in. Only when that happens and the cases slow down on their own can it consider a more extensive reopening like a City A region.

Testing is an important part of the epidemic management process no doubt since it enables the authorities to get an accurate description of the spread of the disease. As we have already discussed, limited testing capacity is giving us a partial or distorted picture in many regions. There is a widespread media perception that extensive testing is one of the prerequisites for any kind of reopening process [7], [8]. Much criticism has also been levelled at certain countries for having inadequate testing programs (we shall further elaborate the blame aspects later). However, we would like to emphasize that testing is as of yet a diagnostic tool and not a preventive one. Currently, it can show us how the disease is behaving but cannot slow its spread in any way. Test-induced slowing can come only when the capacity expands to such a level as to be able to preventively test potential super-spreaders such as grocers and food workers every single day. We hope that such a development may prove possible in the near future – many Universities for example are making reopening arrangements with provision for very frequent testing of the entire community.

During reopening it is vital to get a true picture of the disease evolution so that we can gauge the effect of any relaxation of restrictions – whether it keeps the outbreak under control as in City E or brings about the beginnings of a second wave as in City F. Such beginnings are heralded by a rise in the case rate. As we saw, there was no such rise in City E even though R increased after the reopening. If the rise takes place, the relaxation must immediately be rolled back to avert the disaster. Hence, during reopening, the testing capacity must be high enough to detect such incipient rises. As per China's state media reports, with an aim to reopen the economy, the city of Wuhan conducted 6 million tests in one week; we present this fact without discussion or comment. A second reason why testing is still not all that it could have been is the high false-negative rate during the initial stages of infection [9]. Suppose a contact tracing drive identifies Mr X as a potential case, having been exposed to a known case yesterday. Then, it can be that Mr X contracts the virus ten days from now, in which situation he will report negative if tested today or tomorrow, but will still amount to a spreading risk ten days later if he is at large then. This also means that secondary contact tracing, i.e. finding Mr. X's contacts, must go ahead irrespective of his test results. Indeed, the medical authorities are well aware of this loophole.

The US Chamber of Commerce has given out state by state reopening guides for small businesses which are mandated to be followed across the US. Continued following of federal, state, tribal, territorial and local recommendations is of paramount importance.

Prior to resuming work, all workplaces should have a carefully chartered exposure control, mitigation and recovery plan. Although essential guidance is specific for each business, there are certain measures that can be generally adopted across all workplaces.

1) Reopening in phases – The US government has laid down guidelines to open the country in 3 phases. First phase involves continuation of vulnerable individuals to remain at home. When in public, people are expected to wear masks, have maximum physical separation, avoid places with more than 10 people and limit non-essential travel. Second phase allows gatherings of 50 people, some nonessential travel and reopening of schools. Third phase involves relaxation of restrictions, permitting vulnerable populations to operate.

2) Defining new metrics – Post-corona world will witness some significant changes in regulatory controls, and behavioral drift in personal and professional spheres. Cleanliness standards, safety standards, infection prevention practices with regular monitoring and inspection for its assurance are some of the new terms that will have to be a part of a daily life of the people for at least the next few months.

3) Organizational changes – To help essential operations to function, companies and organizations will have to be prepared with advanced IT systems (in case of continuation of remote working), supply of PPE, setting up travel facilities to avoid public transport, providing behavioral health services, and leave no stone unturned in overcoming biological, physical, and emotional challenges. We can see that the above guidelines are broadly conformal to our model predictions.

4 METHODS OF CONTACT TRACING

As we have already mentioned, contact tracing is probably the single most important factor in determining the progression of COVID-19 in a region. We can see from the model that the faster the contact tracing takes place, the better; the more delay we have, the higher R becomes. Moreover, our model does not account for backward contact tracing. In practice however, a sufficiently high level of detection might not be possible to achieve with forward contact tracing alone. As much as it is important, contact tracing is also one of the trickiest aspects to handle since it can interfere with people's privacy. In classical

contact tracing, human tracers talk to the confirmed cases and track down their movements as well as the persons they interacted with over the past couple of days. This method has worked well in Ithaca, USA and in Kerala, India. While it is the least invasive of privacy, it is also the most unreliable since people might not remember their movements or their interactions correctly. The time taken in this method is also the maximum. A more sophisticated variant of this supplements human testimony with CCTV footage and credit/debit card transaction histories – this approach is possible only in countries such as USA where card usage predominates over cash. The most sophisticated contact tracing algorithms use artificial intelligence together with location-tracking mobile devices and apps – while they are quick and fool-proof, they automatically raise issues of privacy and security. For example, the TraceTogether app in Singapore, which worked very well during the initial phases of the outbreak, has not found popularity with many users [10]. Similarly, India's Aarogya Setu has also raised privacy concerns [11]. Americans too have expressed their aversion to using contact tracing apps in a recent poll, with only 43 percent of people saying that they trusted companies like Google or Apple with their data.

5 ENSURING SOCIAL COMPLIANCE – A BEHAVIORAL PERSPECTIVE

As the epidemic drags on and on, the continued restrictions on social activity are becoming more and more unbearable. There is an increasing tendency, especially among younger people who are much less at risk of serious symptoms, to violate the restrictions and spread the disease through irresponsible actions. However, City F, a rise in violator behavior can completely nullify the effects of lockdown over the past few weeks or months. Here we discuss how public health professionals and policy makers can resort to behavior/psychological theories to ensure compliance among the common people. The most widely used model is Health Belief Model which has been used successfully in addressing public health challenges. We briefly discuss the utility of this model in the current situation.

Health belief model is a theoretical model which hypothesizes that interventions will be most effective if they target key factors that influence health behaviors such as perceived susceptibility, perceived severity, perceived benefits, perceived barriers to action and exposure to factors that prompt action and self-efficacy. In general, this model can be used to design short and long term interventions. The prime components of this model which are relevant in the current scenario can be outlined as follows.

1) Conducting a health need assessment to determine the target population – The best example is the demarcation of zones in India depending on the level of risk. Red zone is highest risk, orange zone is average risk and green zone translates into no cases since last 21 days. Classification is multifactorial, taking into account the incidence of cases, the doubling rate and the limit of testing and surveillance feedback to classify the districts.

2) Communicating the consequences involved with risky behaviors in a transparent manner – Central and state ministers as well as public health authorities are in constant communication with the masses.

3) Conveying information about the steps involved in performing the recommended action and focusing on the benefits to action – Famous celebrities, in addition to state and central governments, spread the messages explaining the required steps cogently and ensuring that it has the maximum reach, especially among social media-addicted millennials and similar populations.

4) Being open about the issues/barriers, identifying them at early stage and working toward resolution – Activating all sorts of helpline numbers, email addresses, personal offices etc to address any grievances around the topic.

5) Developing skills and providing assistance that encourages self-efficacy and possibility of positive behavior change – Adequate arrangements for people from lower socio-economic strata, stable and trustworthy financial schemes for middle class, plan to support small business and a means to become a bridge between the affluent class and the needy class are some of the ways to foster positive behavior change and develop natural trust. Other than health belief model, some theories that can be useful are:

Theory of Reasoned Action – This theory implies that an individual's behavior is based on the outcomes which the individual expects as a result of such behavior. In a practical scenario, if the health officials want the people to follow a particular trend, let us say based on our model, they need to reinforce the advantages of targeted behavior and strategically address the barriers. For instance, to enforce separation minima even when it is apparently proving ineffective and the cases are increasing, they can use the examples of Cities B and C to convince the citizens that violations – and hence violators – can be responsible for thousands of excess deaths. Trans-theoretical Model – This model posits that any health behavior change entails progress through six stages of change: precontemplation, contemplation, preparation, action, maintenance and termination. For instance, it was observed that in March, despite a rise in cases in New York City (NYC), people were not observing social restrictions the way they should have. Now, we can see that with passing time, the behavior of the masses transforms according to the stages of this model

Precontemplation – This is a stage where people are typically not cognizant of the fact that their behavior is troublesome and may cause undesirable consequences. There is a long way to go before an actual behavior change. This phase coincides with the commencement of cases in NYC.

Contemplation – Recognition of the behavior as problematic begins to surface and a shift begins towards behavior change. When the cases started being reported all over media and the major cause of spread began to surface, citizens started paying attention to their activities.

Preparation – People start taking small steps toward behavior change like in our case, exhibiting hygienic practices and ensuring six feet separation minima.

Action – This stage covers the phase where people have just changed their behavior and have positive intention to maintain that approach. In this instance, people continue to practice social restrictions and hygiene positively.

Maintenance – This stage focuses on maintenance and continuity toward the adopted approach. Majority of people in NYC are exhibiting positive behavior and maintaining it throughout the stages of reopening phases. This is vitally important to ensure that NYC stops at partial herd immunity like City D instead of blowing up again like City C.

Termination – There is lack of motivation to come back to the unhealthy behaviors and some sections of people across the country/world will continue practicing good hygiene (though not social restrictions!) in our day-to-day lives.

Social Ecological theory – This theory highlights multiple levels of influences that molds the decision. In our case, let us say for example that the decision is to maintain sufficient physical separation once offices are opened up. To successfully follow this, there is a complex interplay between individual, relationship, community and societal factors that comes into action. Law enforcement authorities need to take this into consideration. A group of individuals when motivated by one another to follow the guidelines, builds a good connection within the society, and in turn there is a high probability to build a healthy network within a defined area. A negative interplay at different levels of motivation may in turn, prove disastrous and cause all efforts go down the drain. A perfect illustration of this in the present condition is how various NGO's are working in conjunction with public health authorities to bring about a change at an individual level by door-to-door campaigning. This propels the behavior of even the most potentially recalcitrant population in the most desirable way i.e. wearing masks and gloves, adopting hand hygiene, being cognizant of symptoms arising in any member of the family and following quarantine rules in case of travel from other states.

6 SOCIAL ATTITUDES AND BEHAVIOUR

In this Section we address another important issue related to the Coronavirus. This is that the widely heterogeneous case profiles in different regions have often led to "corona contests" among these regions. Far too often, the residents of better-off regions are seen heaping scorn on worse-hit regions. We have selected a tiny handful of representative media articles, castigating the approaches of India, USA and Sweden, to show the breadth and vitriol of such commentary [12][13][14] [15][16].A feature common to almost all opinion pieces like this is that their authors do not have the slightest knowledge of the issues involved, either epidemiological or economic.

Before embarking on criticisms, we should note that policy decisions need to be taken in real time, as the situation evolves. The authorities do NOT have the benefit of hindsight to decide on their course of action. Since the virus is a new one, there is no precedent which can act as a model. Even among emerging infectious diseases, this latest one is particularly unpredictable, since minuscule changes in parameters can cause dramatic changes in the system's behavior. This phenomenon is best illustrated by the notional cities, discussed previously. For example, to get from City A to B, all we did was increase by 50 percent the fraction of people who escaped the contact-tracers' net. The result was a 30 times (not 30 percent!) increase in the total number of cases. Similarly, the difference between Cities B and D is an 11-day delay (recall that the first seven days in the plots are the seeding period, so they don't count) in imposing the lockdown in D. 11 days out of a 200-plus-day run might not sound like a lot. But, that was enough to create tens of thousands of additional cases, risk overstressing healthcare systems and at the same time shorten the epidemic duration by a factor of three.

Further uncertainty comes from the fact that the parameter values are changing constantly. It is a well- known fact the reported fraction of asymptomatic carriers has increased continuously over the last three months or so. Considering the sensitivity of this or any other model to parameter values, such changes can completely invalidate the results of a model as well as any decision which was made on their basis. Identifying potential exposures is much easier in a smaller city than a large or densely populated one. It is also more effective if the cases are mostly from the sophisticated social class who can use mobile phone contact tracing apps or otherwise keep (at least mental) records of their movements and of the people they interacted with. However, if there is an outbreak among the unsophisticated class, then even the most skillful contact tracer might run up against a wall of zero or false information. In such cases there are limited options that are left to the authorities to proceed in a conducive manner.

India went into lockdown on 25 March 2020. At that time, the official figures stated that there were only 571 cases, which made the decision appear premature to many people. Indeed, a seven-day delay of lockdown was suggested so that the migrant workers would have been able to return to their homes. However, when the lockdown was imposed, the testing had also been woefully inadequate, with a nationwide total of just 22,694 tests having been conducted up to that date. If we use the extrapolation technique of inferring case counts from death counts, then using the same 1 percent mortality rate and 20 day interval to death, we find almost 40,000 assumed cases on the day that the lockdown began. If we go by this figure, then the lockdown wasn't really early, and possibly should have been enforced earlier still in trouble zones such as Mumbai. Certainly, if the figure of 40,000 cases is true, then one further week of normal life (with huge crowds in trains and railway stations) might have been disastrous. From the vantage point of today, alternate arrangements should definitely have been made much earlier for rehabilitation of the migrant workers. However these arrangements would have involved considerable complexity in the prevailing situation, and were certainly not as easy as one

week's delay in announcing lockdown. Sweden, which has adopted a controlled herd immunity strategy, has been accused of playing with fire. It is also possible that the Swedish authorities are aware that they do not have the contact tracing capacity required for performing like City A and hence are attempting something like City D – a faster end of the epidemic than City B at the expense of a higher case count. To make a comprehensive analysis of their policy, it is crucial to know not only the last intricate detail of the epidemiological aspects but also the details of the economic considerations. That is almost impossible. On a different note however, we have seen reports [17], [18] stating that the virus has entered into old age homes and similar establishments, causing hundreds of deaths over there. Assuming that these reports are not overturned in the course of time, allowing the ingress of virus into high-risk areas is an indefensible action, whatever the overall epidemiological strategy.

Finally, extremely important public health factors such as the racial dependence of susceptibility and/or transmissibility have just started coming to the surface. Another complete grey area is the mutations which this new and vicious virus are undergoing and what effect they might have on the spreading dynamics. Some reports also reflect that the change in genetic composition due to mutation might be the reason behind huge differences in the crude infection rate between countries [19][20]. In the absence of a clear picture about this, any public health measure is all the more likely to be a random guess with non-zero probabilities of both success and failure. Not everything about corona is random or outside one's control though. Amongst the European countries, we can see that Germany, Austria, Switzerland, Denmark, Norway and Finland have definitely managed the epidemic while their neighbors have not, which rules out some hidden luck factor. The same has happened in Kerala and Karnataka (also in India). This has been feasible only due to governmental awareness and hard work, and people's cooperation. Similarly, there are some governments which have been clearly guilty of negligence or hubris in their management of the disease. It would also be noteworthy to observe and take lessons from the some of the new places like Alabama, Arkansas, Florida , Texas etc which have been recently identified as potential hotspots of this pandemic. Lastly, our conclusion best resonates with the message that coronavirus is not some kind of race but a public health disaster and we should adopt a unified approach to the fight against it.

CONCLUSION

Here, we summarize the take-home messages from this paper:

- A city can reopen only if it is past the peak of cases. Reopening must be accompanied by robust contact tracing. The US CDC has laid down a set of reopening guidelines which are compatible with our model and its solutions.

- Incorporation of socio-behavioral theories can come into play for effective execution of interventional strategies.

- Efficiency of contact tracing comes at the expense of people's privacy – balancing between the two is a delicate optimization problem.

- In some regions, restrictions such as masks and six-feet separation minima must be maintained for a very long time to come. The public health authorities can ensure compliance by resorting to socio –behavioral theories/approaches.

- In deploying advanced contact tracing techniques, significant consideration has to be given for ensuring high data security and lay down privacy regulations that are convincing to the users

- Control the spread by swift identification and isolation of cases accompanied by tracing and quarantine for at least 2 weeks

- Empowering of individuals and communities by the government to facilitate efficient capacity building.

- Multidisciplinary coordination, strong leadership to mobilize communities and take quick decisions coupled with thoughtful development of operation plans are likely to prove considerably efficient in handling this pandemic to the best of our capacity.

References

[1] B. Shayak and M. M. Sharma, "Retarded logistic equation as a universal dynamic model for the spread of COVID-19," *medRxiv*, p. 2020.06.09.20126573, 2020, doi: 10.1101/2020.06.09.20126573.

[2] E. Okanyene, B. Rader, Y. L. Barnoon, L. Goodwin, and J. S. Brownstein, "Analysis of hospital traffic and search engine data in Wuhan China indicates early disease activity in the Fall of 2019," *Harvard*, 2020, [Online]. Available: http://nrs.harvard.edu/urn-3:HUL.InstRepos:42669767.

[3] CDC, "Forecasting COVID-19 in the US," 2020. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html.

[4] "Microsoft coronavirus webpage." https://www.bing.com/covid.

[5] "COVID-19 in India." [Internet]. Available from: https://www.covid19india.org/.

[6] L. Star and S. Moghadas, "The Role of Mathematical Modelling in Public Health Planning and Decision Making," *Natl. Collab. Cent. Infect. Dis.*, vol. (5)2, no. 2, pp. 285–299, 2010.

[7] Livemint, ""Many states are far short of COVID-19 testing levels." https://www.statnews.com/2020/04/27/coronavirus-many-states-short-of-testing-levels-needed-for-safereopening/.

[8] Harvard Business Review, “A Plan to Safely Reopen the U.S. Despite Inadequate Testing.” https://hbr.org/2020/05/a-plan-to-safely-reopen-the-u-s-despite-inadequate-testing.

[9] S. Telles, S. K. Reddy, and H. R. Nagendra, “Variation in False Negative Rate of RT-PCR Based SARS-CoV-2 Tests by Time Since Exposure,” *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2019, doi: 10.1017/CBO9781107415324.004.

[10] M. Lee, “Given low adoption rate of TraceTogether, experts suggest merging with SafeEntry or other apps,” *Today*, 2020. https://www.todayonline.com/singapore/given-low-adoption-rate-tracetogether-experts-suggest-merging-safeentry-or-other-apps.

[11] A. Zargar, “Privacy, security concerns as India forces virus-tracking app on millions,” *CBS News*. .

[12] K. Bajpai, “Five lessons of COVID.” Available: from: https://timesofindia.indiatimes.com/blogs/toi-editpage/five-lessons-of-covid-factors-that-are-negative-for-india-are-having-greater-impact-than-mitigating-ones/..

[13] K. Grimes, “Is politics the reason why Gov. Newsom is keeping California locked down ?,” *California Globe*. .

[14] R.Guha, “What Modi got wrong on COVID-19 and how he can fix it.” https://www.ndtv.com/opinion/5-lessons-for-modi-on-covid-19-by-ramachandra-guha-2227259.

[15] K. Weintraub, “Sweden sticks with controverial covid approach.,” [Online]. Available: https://www.webmd.com/lung/news/20200501/sweden-sticks-with-controversial-covid19-approach.

[16] The Island Now, “Cuomo has failed in his handling of coronavirus.” https://theislandnow.com/opinions-100/readers-write-cuomo-has-failed-in-handling-of-coronavirus/.

[17] “Are care homes the dark side of Sweden’s coronavirus strategy.” https://www.euronews.com/2020/05/19/are-care-homes-the-dark-side-of-sweden-s-coronavirus-strategy.

[18] “What’s going wrong in Sweden’s care homes.” .

[19] L. van Dorp *et al.*, “Emergence of genomic diversity and recurrent mutations in SARS-CoV-2,” *Infect. Genet. Evol.*, vol. 83, no. May, p. 104351, 2020, doi: 10.1016/j.meegid.2020.104351.

[20] H. Ellyatt, “Coronavirus no longer exists clinically - controversy,” *CNBC*. https://www.cnbc.com/2020/06/02/claim-coronavirus-no-longer-exists-provokes-controversy.html.

Attention Realignment and Pseudo-Labelling for Interpretable Cross-Lingual Classification of Crisis Tweets

Jitin Krishnan
Department of Computer Science
George Mason University
Fairfax, VA
jkrishn2@gmu.edu

Hemant Purohit
Department of Information Sciences & Technology
George Mason University
Fairfax, VA
hpurohit@gmu.edu

Huzefa Rangwala
Department of Computer Science
George Mason University
Fairfax, VA
rangwala@gmu.edu

ABSTRACT

State-of-the-art models for cross-lingual language understanding such as XLM-R [7] have shown great performance on benchmark data sets. However, they typically require some fine-tuning or customization to adapt to downstream NLP tasks for a domain. In this work, we study unsupervised cross-lingual text classification task in the context of crisis domain, where rapidly filtering relevant data regardless of language is critical to improve situational awareness of emergency services. Specifically, we address two research questions: a) Can a custom neural network model over XLM-R trained only in English for such classification task transfer knowledge to multilingual data and vice-versa? b) By employing an attention mechanism, does the model attend to words relevant to the task regardless of the language? To this goal, we present an attention realignment mechanism that utilizes a parallel language classifier to minimize any linguistic differences between the source and target languages. Additionally, we pseudo-label the tweets from the target language which is then augmented with the tweets in the source language for retraining the model. We conduct experiments using Twitter posts (tweets) labelled as a 'request' in the open source data set by Appen[1], consisting of multilingual tweets for crisis response. Experimental results show that attention realignment and pseudo-labelling improve the performance of unsupervised cross-lingual classification. We also present an interpretability analysis by evaluating the performance of attention layers on original versus translated messages.

KEYWORDS

Social Media, Crisis Management, Text Classification, Unsupervised Cross-Lingual Adaptation, Interpretability

ACM Reference Format:
Jitin Krishnan, Hemant Purohit, and Huzefa Rangwala. 2020. Attention Realignment and Pseudo-Labelling for Interpretable Cross-Lingual Classification of Crisis Tweets. In *Proceedings of KDD Workshop on Knowledge-infused Mining and Learning (KiML'20).* , 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1]https://appen.com/datasets/combined-disaster-response-data/

KiML'20, August 24, 2020, San Diego, California, USA,

https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Social media platforms such as Twitter provide valuable information to aid emergency response organizations in gaining real-time situational awareness during the sudden onset of crisis situations [4]. Extracting critical information about affected individuals, infrastructure damage, medical emergencies, or food and shelter needs can help emergency managers make time-critical decisions and allocate resources efficiently [15, 21, 22, 30, 31, 36]. Researchers have designed numerous classification models to help towards this humanitarian goal of converting real-time social media streams into actionable knowledge [1, 22, 26, 28, 29]. Recently, with the advent of multilingual models such as multilingual BERT [9] and XLM [20], researchers have started adopting them to multilingual disaster tweets [6, 25]. Since XLM-R [7] has been shown to be the most superior model in cross-lingual language understanding, we restrict our work to this model to explore the aspects of cross-lingual transfer of knowledge and interpretability.

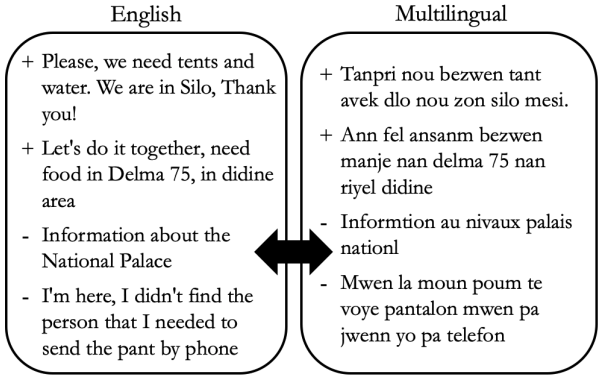


Figure 1: Problem: Unsupervised cross-lingual tweet classification, e.g., train a model using English tweets, predict labels for Multilingual tweets, and vice-versa.

In this work, we address two questions. First is to examine whether XLM-R is effective in capturing multilingual knowledge by constructing a custom model over it to analyze if a model trained using English-only tweets will generalize to multilingual data and vice-versa. Social media streams are generally different from other text, given the user-generated content. For example, tweets are usually short with possibly errors and ambiguity in the behavioral expressions. These properties in turn make the classification task or extracting representations a bit more challenging. Second question

is to examine whether word translations will be equally attended by the attention layers. For instance, the words with higher attention weights in a sentence in Haitian Creole such as "*Tanpri nou bezwen tant avek dlo nou zon silo mesi*" should align with the words in its corresponding translated tweet in English "*Please, we need tents and water. We are in Silo, Thank you!*". Our core idea is that if '*dlo*' in the Haitian tweet has a higher weight, so should its English translation '*water*'. This word-level language agnostic property can promote machine learning models to be more interpretable. This also brings several benefits to downstream tasks such as knowledge graph construction using keywords extracted from tweets. In situations where data is available only in one language, this similarity in attention would still allow us to extract relevant phrases in cross-lingual settings. To the best of our knowledge in crisis analytics domain, aligning attention in cross-lingual setting is not attempted before. In this work, we focus our classification experiments only to tweets containing '*request*' intent, which will be expanded to other behaviors, tasks, and datasets in the future.

Contributions: We propose a novel attention realignment method which promotes the task classifier to be more language agnostic, which in turn tests the effectiveness of multilingual knowledge capture of XLM-R model for crisis tweets; and a pseudo-labelling procedure to further enhance the model's generalizability. Furher, incorporating the attention-based mechanism allows us to perform an interpretability analysis on the model, by comparing how words are attended in the original versus translated tweets.

2 RELATED WORK AND BACKGROUND

There are numerous prior works (*c.f.* surveys [4, 14]) that focus specifically on disaster related data to perform classification and other rapid assessments during an onset of a new disaster event. Crisis period is an important but challenging situation, where collecting labeled data during an ongoing event is very expensive. This problem led to several works on domain adaptation techniques in which machine learning models can learn and generalize to unseen crisis event [3, 10, 18, 23]. In the context of crisis data, Nguyen et al. [28] designed a convolutional neural network model which does not require any feature engineering and Alam et al. [1] designed a CNN architecture with adversarial training on graph embeddings. Krishnan et al. [19] showed that sharing a common layer for multiple tasks can improve performance of tasks with limited labels.

In multilingual or cross-lingual direction, many works [8, 17] tried to align word embeddings (such as fastText [27]) from different languages into the same space so that a word and its translations have the same vector. These models are superseded by models such as multilingual BERT [9] and XLM-R [7] that produce contextual embeddings which can be pretrained using several languages together to achieve impressive performance gains on multilingual use-cases.

Attention mechanism [2, 24] is one of the most widely used methods in deep learning that can construct a context vector by weighing on the entire input sequence which improves over previous sequence-to-sequence models [13, 34, 35]. As the model produces weights associated with each word in a sentence, this allows for evaluating interpretability by comparing the words that are given priority in original versus translated tweets.

With more and more machine learning systems being adopted by diverse application domains, transparency in decision-making inevitably becomes an essential criteria, especially in high-risk scenarios [12] where trust is of utmost importance. With deep neural networks, including natural language systems, shown to be easily fooled [16], there has been many promising ideas that empower machine learning systems with the ability to explain their predictions [5, 32]. Gilpin et al. [11] presents a survey of interpretability in machine learning, which provides a taxonomy of research that addresses various aspects of this problem. Similar to the work by Ross et al. [33], we employ an attention-based approach to evaluate model interpretability applied to the crisis-domain.

3 METHODOLOGY

3.1 Problem Statement: Unsupervised Cross-Lingual Crisis Tweet Classification

Consider tweets in language A and their corresponding translated tweets in language B. The task of unsupervised cross-lingual classification is to train a classifier using the data only from the source language and predict the labels for the data in the target language. This experimental set up is usually represented as $A \rightarrow B$ for training a model using A and testing on B or $A \rightarrow B$ for training a model using B and testing on A. X refers to the data and Y refers to the ground truth labels. The multilingual dataset used in our experiments consists of original multilingual (*ml*) tweets and their translated (*en*) tweets in English. To summarize:

Experiment A ($en \rightarrow ml$):
Input: X_{en}, Y_{en}, X_{ml}
Output: $Y_{ml}^{pred} \leftarrow predict(X_{ml})$
Experiment B ($ml \rightarrow en$):
Input: X_{ml}, Y_{ml}, X_{en}
Output: $Y_{en}^{pred} \leftarrow predict(X_{en})$

3.2 Overview

In the following sections, we propose two methodologies to enhance cross-lingual classification: 1) Attention Realignment and 2) Pseudo-Labelling. Attention realignment utilizes a language classifier which is trained in parallel to realign the attention layer of the task classifier such that the weights are more geared towards task-specific words regardless of the language. Pseudo-Labelling further enhances the classifier by adding high quality seeds from the target language that are pseudo-labelled by the task classifier.

3.3 Attention Realignment by Parallel Language Classifier

As depicted in Fig 2, model on the left side is the task classifier and the model on the right side is a language classifier that is trained in parallel. The purpose of this language classifier is to pick up aspects that is missed by the XLM-R model. This could be tweet-specific, crisis-specific, or other linguistic nuances that can separate original tweets and translated tweets. Note that semantically, translated words are expected to have similar XLM-R representations.

Attention realignment is a mechanism we introduce to promote the task classifier to be more language independent. The main idea is that the words that are given higher attention in a language

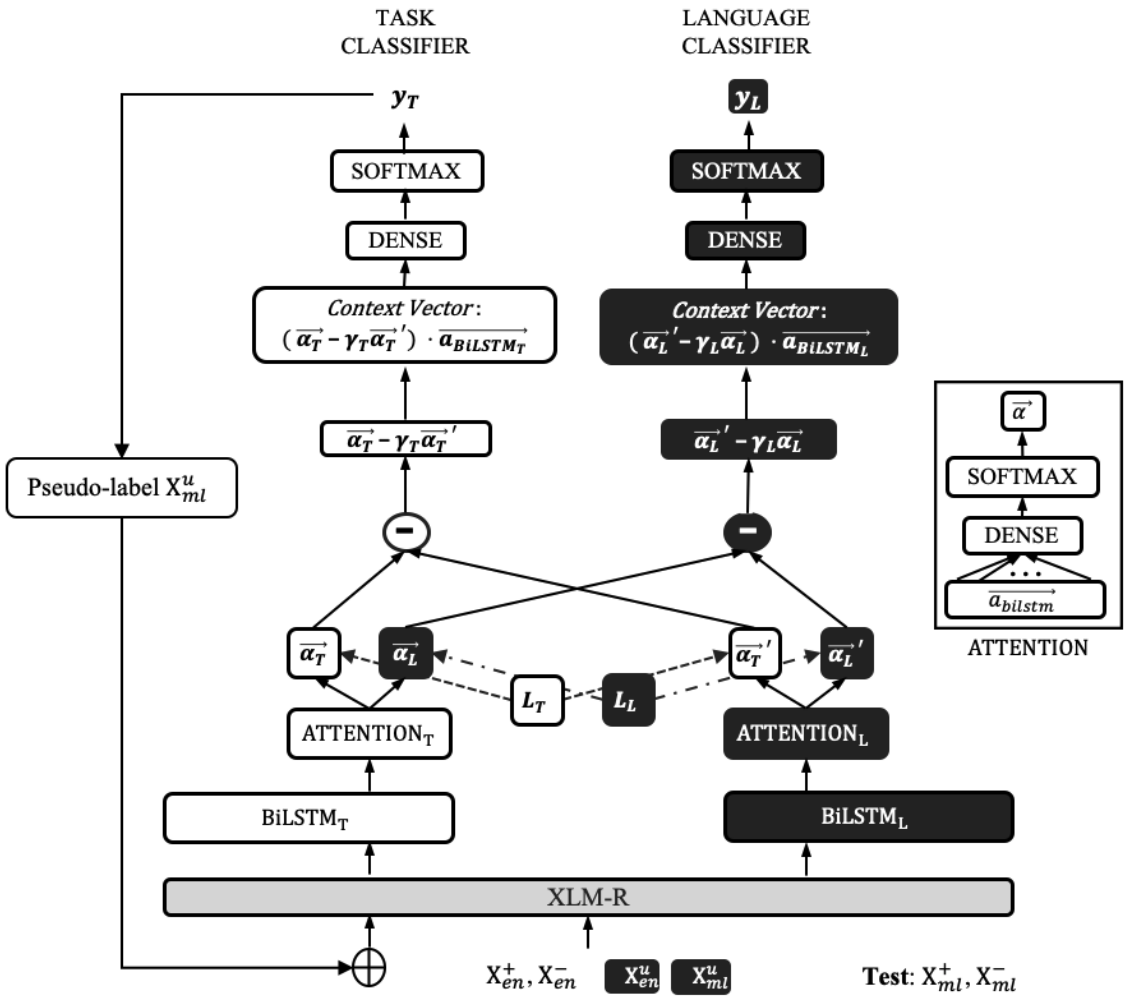


Figure 2: Attention Realignment with Pseudo-Labelling over XLM-R model

Notation	Definition
en	Tweets translated to English ('message' column in the dataset)
ml	Multilingual Tweets ('original' column in the dataset)
α	Attention Layer
T	A component that uses Task-specific data. i.e., + and − 'Request' tweets
L	A component that uses Language-specific data. i.e., *en* and *ml* tweets
a_{BiLSTM}	Activation from the BiLSTM layer
β, γ, ζ	Hyperparameters

Table 1: Notations

classifier should be less important in a task classifier. For example, '*dlo*' in Haitian and '*water*' in English should have the same vector representation in language agnostic models; while the sentence structure, grammar, and other nuances can vary. We enforce this rule by constructing two operations:

(1) **Attention Difference**: When a sentence goes through model M1, it also goes through model M2. For the same sentence, this returns two attention layer weights: one from the task classifier ($\overrightarrow{\alpha_T}$) and the other from the language classifier ($\overrightarrow{\alpha_T}'$). Directly subtracting $\overrightarrow{\alpha_T}'$ from $\overrightarrow{\alpha_T}$ poses two issues: 1) we do not know whether they are comparable and 2) $\overrightarrow{\alpha_T}'$ may have negative values. A simple solution to this is to normalize bothe vectors and clip $\overrightarrow{\alpha_T}'$ such that it is between 0 and 1. Thus, an attention subtraction step is as follows:

$$\frac{\overrightarrow{\alpha_T}}{\|\overrightarrow{\alpha_T}\|} - \gamma_T \, clip\left(\frac{\overrightarrow{\alpha_T}'}{\|\overrightarrow{\alpha_T}'\|}, 0, 1\right) \tag{1}$$

where γ_T is a hyperparameter to tune the amount of subtraction needed for the task classifier. Similarly, for the language

classifier,

$$\frac{\overrightarrow{\alpha_L}'}{\|\overrightarrow{\alpha_L}'\|} - \gamma_L \, clip\left(\frac{\overrightarrow{\alpha_L}}{\|\overrightarrow{\alpha_L}\|}, 0, 1\right) \tag{2}$$

(2) **Attention Loss**: Along with attention difference, the model can also be trained by inserting an additional loss function term that penalizes the similarity between the attention weights from the two classifiers. We use the Frobenius norm.

$$L_{At} = \|\overrightarrow{\alpha_T}^T \, \overrightarrow{\alpha_T}'\|_F^2 \tag{3}$$

$$L_{Al} = \|\overrightarrow{\alpha_L}^T \, \overrightarrow{\alpha_L}'\|_F^2 \tag{4}$$

for task and language respectively. Resulting final loss function of joint training will be:

$$L(\theta) = \zeta_T\Big(CE_T + \beta_T L_{At}\Big) + \zeta_L\Big(CE_L + \beta_L L_{Al}\Big) \tag{5}$$

where β is the hyperparameter to tune the attention loss weight, ζ is the hyperparameter to tune the joint training loss, and CE denotes the binary cross entropy loss,

$$CE = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)] \tag{6}$$

It is important to note that the Frobenius norm is **not** simply between the attention weights of the two models but rather between the attention weights produced by the two models on the same input tweet. For example, for a given tweet, the task classifier attends more to task-specific words and the language classifier attends to language-specific words. So the mechanism makes sure that they are distinct.

3.4 Pseudo-Labelling

To enhance the model further, we pseudo-label the data in the target language. For example, if we are training a model using the English tweets, we use the original tweets before translation for pseudo-labelling. The idea is simply to gather high-quality seeds from the target to retrain the model. Note that, we still do not use any target labels here; still following the unsupervised goal. Thus, for retraining model M1 for $en \rightarrow ml$, the new dataset would consist of: X_{en}^{+} and $X_{ml}^{pseudo+}$ as positive examples and X_{en}^{-} and $X_{ml}^{pseudo-}$ as negative examples.

3.5 XLM-R Usage

The recommended feature usage of XLM-R[2] is either by fine-tuning to the task or by aggregating features from all the 25 layers. We employ the later to extract the multilingual embeddings for the tweets.

4 DATASET & EXPERIMENTAL SETUP

	Train	Validation	Test
Positive	3554	418	496
Negative	17473	2152	2128

Table 2: Dataset Statistics for both *en* amd *ml*

[2]https://github.com/facebookresearch/XLM

T_x	30
Deep Learning Library	Keras
Optimizer	Adam [$lr = 0.005$, $beta_1 = 0.9$, $beta_2 = 0.999$, $decay = 0.01$]
Maximum Epoch	100
Dropout	0.2
Early Stopping Patience	10
Batch Size	32
ζ_T	1
ζ_L	0.1
$\beta_T, \beta_L, \gamma_T, \gamma_L$	0.01

Table 3: Implementation Details

We use the open source dataset from Appen[3] consisting of multilingual crisis response tweets. The dataset statistics for tweets with 'request' behavior labels is shown in Table 2. For all the experiments, the dataset is balanced for each split.

Each experiment is denoted as $A \rightarrow B$, where A is the data that is used to train the model and B is the data that is used for testing the model. For example, $en \rightarrow ml$ means we train the model using English tweets and test on multilingual tweets.

Models are implemented in Keras and the details are shown in table 3. Hyperparameters β_T, β_L, γ_T, and γ_L are not exhaustively tuned; we leave this exploration for future work.

	Baseline	Model M1	Model M2
$en \rightarrow ml$	59.98 (80.57)	62.53 (77.02)	**66.79** (82.39)
$ml \rightarrow en$	60.93 (70.07)	65.69 (63.50)	**70.95** (73.84)

Table 4: Performance Comparison (Accuracy in %) for *Source* → *Target* (*Source* → *Source*).
Baseline = XLMR + BiLSTM + Attention.
Model M1 = Baseline + Attention Realignment.
Model M2 = Model M1 + Pseudo-Labelling.

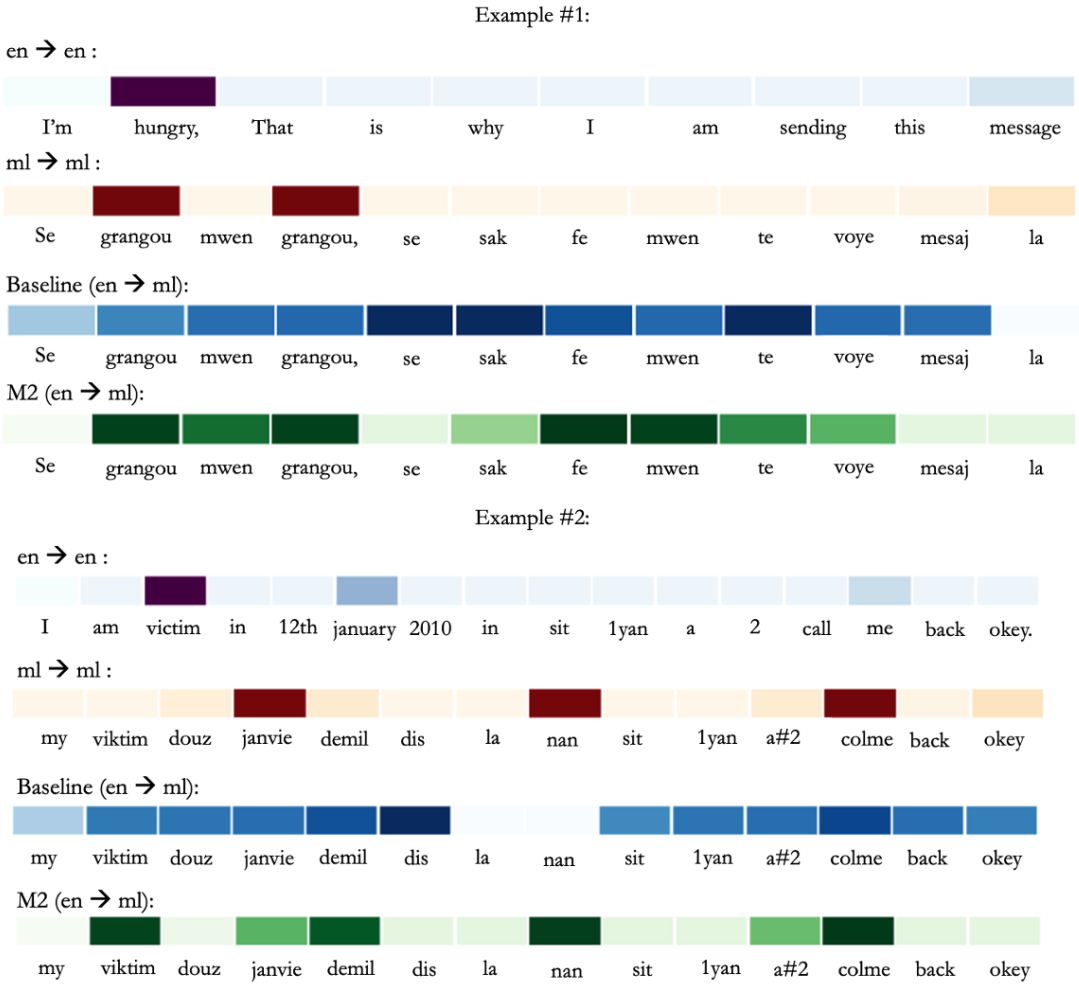
5 RESULTS & DISCUSSION

Table 4 shows the cross-lingual performance comparison of all the models. The three models are described below:

(1) **Baseline**: The baseline model consists of embeddings retrieved from XLM-R trained over BiLSTMs and Attention layers. This is a traditional sequence (text) classifier enhanced with attention mechanism. Activations from the BiLSTM layers are weighed by the attention layer to construct the context vector which is then passed through a dense layer and softmax function to produce the classification output.
(2) **Model M1**: Adding attention realignment to the baseline model produces model M1. Attention realignment is achieved through a language classifier which is trained in parallel with the goal to make the task classifier more language agnostic.

[3]https://appen.com/datasets/combined-disaster-response-data/

Figure 3: Attention visualization example for '*request*' tweets: words and their attention weights for two tweets in Haitian Creole and its translation in English (darker the shade, higher the attention).



The attention weights for both task and language classifiers are manipulated by each other during training by a process of subtraction (attention difference) as well a loss component (attention loss). See section 3.3.

(3) **Model M2**: Adding the pseudo-labelling procedure to model M1 produces model M2. Using Model M1 which is trained to be language agnostic, tweets from the target languages are pseudo-labelled. High quality seeds are selected (using Model M1 p>0.7) and augmented to the original training dataset to retrain the task classifier.

Results show that, for cross-lingual evaluation on $en \rightarrow ml$, model M1 outperforms the baseline by **+4.3%** and model M2 outperforms by **+11.4%**. On $ml \rightarrow en$, model M1 outperforms the baseline by **+7.8%** and model M2 outperforms by **+16.5%**. This shows that both models are effective in cross-lingual crisis tweet classification. An interesting observation to note is that using attention realignment alone decreased the classification performance in the same language, which is brought back up by pseudo-labelling. These scores are shown in brackets in table 4. A deeper investigation in this direction on various other tasks can shed more light on the impact of realignment mechanism.

5.1 Interpretability: Attention Visualization

We follow a similar attention architecture shown in [18]. The context vector is constructed as a result of dot product between the attention weights and word activations. This represents the interpretable layer in our architecture. The attention weights represent the importance of each word in the classification process. Two examples are shown in figure 3. In the first example, both $en \rightarrow en$ and $ml \rightarrow ml$ give attention to the word '*hungry*' (i.e., '*grangou*' in Haitian Creole). Note that these two are results from the models that are trained in the same language in which they are tested; thus, expecting an ideal performance. For the baseline model in the cross-lingual set-up $en \rightarrow ml$, although it correctly predicts the label, the attention weights are more spread apart. In model M2 with attention realignment and pseudo-labelling, although with some spread,

the attention weights are shifted more toward '*grangou*'. Similarly in example 2, the attention weights in the baseline model are more spread apart. Cross-lingual performance of model M2 aligns more with $en \rightarrow en$ and $ml \rightarrow ml$. These examples show the importance of having interpretability as a key criterion in cross-lingual crisis tweet classification problems; which can also be used for downstream tasks such as extracting relevant keywords for knowledge graph construction.

6 CONCLUSION

We presented a novel approach for unsupervised cross-lingual crisis tweet classification problem using a combination of attention realignment mechanism and a pseudo-labelling procedure (over the state-of-the-art multilingual model XLM-R) to promote the task classifier to be more language agnostic. Performance evaluation showed that both models M1 and M2 outperformed the baseline by +4.3% and +11.4% respectively for cross-lingual text classification from English to Multilingual. We also presented an interpretability analysis by comparing the attention layers of the models. It shows the importance of incorporating a word-level language agnostic characteristic in the learning process, when training data is available only in one language. Performing extensive hyperparameter tuning and expanding the idea to other tasks (including cross-task/multi-task) are left as future work. We also plan another direction for future work as to incorporate the human-engineered knowledge from the multilingual knowledge graphs such as BabelNet in our model architecture that could improve the learning of similar concepts across languages critical to the crisis response agencies.

Reproducibility: Source code is available available at: https://github.com/jitinkrishnan/Cross-Lingual-Crisis-Tweet-Classification

7 ACKNOWLEDGEMENT

Authors would like to thank U.S. National Science Foundation grants IIS-1815459 and IIS-1657379 for partially supporting this research.

REFERENCES

[1] Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018. Domain adaptation with adversarial training and graph embeddings. *arXiv preprint arXiv:1805.05151* (2018).

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[3] John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*. 120–128.

[4] Carlos Castillo. 2016. *Big crisis data: social media in disasters and time-critical situations*. Cambridge University Press.

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*. 2172–2180.

[6] Jishnu Ray Chowdhury, Cornelia Caragea, and Doina Caragea. 2020. Cross-Lingual Disaster-related Multi-label Tweet Classification with Manifold Mixup. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. 292–298.

[7] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

[8] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word Translation Without Parallel Data. *arXiv preprint arXiv:1710.04087* (2017).

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[10] Yaroslav Ganin and Victor Lempitsky. 2014. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495* (2014).

[11] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[12] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2 (2017).

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[14] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)* 47, 4 (2015), 1–38.

[15] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. *arXiv preprint arXiv:1605.05894* (2016).

[16] Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328* (2017).

[17] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. *arXiv preprint arXiv:1804.07745* (2018).

[18] Jitin Krishnan, Hemant Purohit, and Huzefa Rangwala. 2020. Diversity-Based Generalization for Neural Unsupervised Text Classification under Domain Shift. *https://arxiv.org/pdf/2002.10937.pdf* (2020).

[19] Jitin Krishnan, Hemant Purohit, and Huzefa Rangwala. 2020. Unsupervised and Interpretable Domain Adaptation to Rapidly Filter Social Web Data for Emergency Services. *arXiv preprint arXiv:2003.04991* (2020).

[20] Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291* (2019).

[21] Kathy Lee, Ankit Agrawal, and Alok Choudhary. 2013. Real-time disease surveillance using twitter data: demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1474–1477.

[22] Hongmin Li, Doina Caragea, Cornelia Caragea, and Nic Herndon. 2018. Disaster response aided by tweet classification with a domain adaptation approach. *Journal of Contingencies and Crisis Management* 26, 1 (2018), 16–27.

[23] Zheng Li, Ying Wei, Yu Zhang, and Qiang Yang. 2018. Hierarchical attention transfer network for cross-domain sentiment classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[24] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[25] Guoqin Ma. 2019. Tweets Classification with BERT in the Field of Disaster Management. https://pdfs.semanticscholar.org/d226/185fa1e14118d746cf0b04dc5be8f545ec24.pdf.

[26] Reza Mazloom, Hongmin Li, Doina Caragea, Cornelia Caragea, and Muhammad Imran. 2019. A Hybrid Domain Adaptation Approach for Identifying Crisis-Relevant Tweets. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)* 11, 2 (2019), 1–19.

[27] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[28] Dat Tien Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. 2016. Rapid classification of crisis-related data on social networks using convolutional neural networks. *arXiv preprint arXiv:1608.03902* (2016).

[29] Ferda Ofli, Patrick Meier, Muhammad Imran, Carlos Castillo, Devis Tuia, Nicolas Rey, Julien Briant, Pauline Millet, Friedrich Reinhard, Matthew Parkan, et al. 2016. Combining human computing and machine learning to make sense of big (aerial) data for disaster response. *Big data* 4, 1 (2016), 47–59.

[30] Bahman Pedrood and Hemant Purohit. 2018. Mining help intent on twitter during disasters via transfer learning with sparse coding. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 141–153.

[31] Hemant Purohit, Carlos Castillo, Fernando Diaz, Amit Sheth, and Patrick Meier. 2013. Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday* 19, 1 (Dec. 2013).

[32] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[33] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717* (2017).
[34] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45, 11 (1997), 2673–2681.
[35] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.
[36] István Varga, Motoki Sano, Kentaro Torisawa, Chikara Hashimoto, Kiyonori Ohtake, Takao Kawai, Jong-Hoon Oh, and Stijn De Saeger. 2013. Aid is out there: Looking for help from tweets during a large scale disaster. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1619–1629.