Pelajaran dari Arsip: Strategi untuk Mengumpulkan Sosiokultural Data dalam Pembelajaran Mesin

Eun Seo Jo
Universitas Stanford
eunseo@stanford.edu

Timnit Gebru
Google
tgebru@google.com

ABSTRAK

Semakin banyak pekerjaan yang menunjukkan bahwa banyak masalah dalam keadilan, akuntabilitas, transparansi, dan etika dalam sistem pembelajaran mesin berakar pada keputusan seputar pengumpulan data dan proses anotasi. Namun, terlepas dari sifat dasarnya, pengumpulan data tetap menjadi bagian yang diabaikan dari pipeline machine learning (ML). Dalam makalah ini, kami berpendapat bahwa spesialisasi baru harus dibentuk dalam ML yang difokuskan pada metodologi pengumpulan dan anotasi data: upaya yang memerlukan kerangka kerja dan prosedur kelembagaan. Khusus untuk data sosiokultural dapat ditarik kesejajaran dari arsip dan perpustakaan. Arsip adalah upaya komunal terlama untuk mengumpulkan informasi manusia dan pengarsipan telah mengembangkan bahasa dan prosedur untuk mengatasi dan mendiskusikan banyak tantangan yang berkaitan dengan pengumpulan data seperti persetujuan, kekuasaan, inklusivitas, transparansi, dan etika & privasi. Kami membahas lima pendekatan utama dalam praktik pengumpulan dokumen di arsip yang dapat menginformasikan pengumpulan data di bidang sosiokultural

ML. Dengan menunjukkan praktik pengumpulan data dari bidang lain, kami mendorong penelitian ML agar lebih sadar dan sistematis dalam pengumpulan data serta memanfaatkan keahlian lintas disiplin.

KONSEP CCS

• Metodologi komputasi → Pembelajaran mesin.

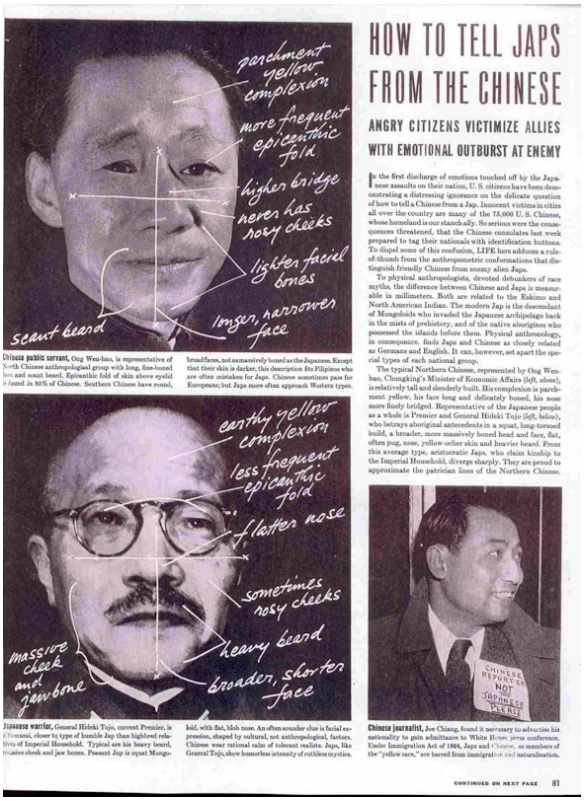
KATA KUNCI

kumpulan data, pembelajaran mesin, keadilan ML, pengumpulan data, data sosial budaya, arsip

Format Referensi ACM:
Eun Seo Jo dan Timnit Gebru. 2020. Pelajaran dari Arsip: Strategi untuk Mengumpulkan Data Sosiokultural dalam Pembelajaran Mesin. Di *Konferensi pada Keadilan, Akuntabilitas, dan Transparansi (FAT * '20), 27-30 Januari 2020, Barcelona, Spanyol.* ACM, New York, NY, AS, 11 halaman. https://doi.org/10. 1145 / 3351095.3372829

1. PERKENALAN

Komposisi data sering kali menentukan hasil penelitian dan sistem pembelajaran mesin (ML). Mengelompokkan orang secara sembarangan dalam data yang digunakan untuk melatih model ML dapat membahayakan kelompok rentan dan menyebarkan bias sosial. Alat otomatis seperti perangkat lunak pengenalan wajah dapat mengekspos kelompok target, terutama dalam kasus daya



Gambar 1: Artikel dari majalah LIFE (Desember 1941) dengan dua gambar yang menunjukkan perbedaan fenotipe yang dapat diidentifikasi antara kelompok ("sekutu") Jepang dan Cina dengan maksud untuk membenci orang Jepang-Amerika setelah pemboman Jepang di Pearl Harbor.

ketidakseimbangan di mana lembaga tertentu memiliki akses eksklusif ke data dan model yang kuat. Secara historis, ciri fenotipe biologis telah digunakan untuk memilih kelompok sasaran pada saat-saat permusuhan publik (Gbr. 1), dan kasus penggunaan serupa telah dilaporkan hari ini dengan teknologi pengenalan wajah [20, 44, 48]. [1] Kasus penggunaan ini menunjukkan para penari dalam membuat kumpulan data besar yang dianotasi dengan ciri fenotipik orang.

Di sisi lain, dalam aplikasi seperti deteksi melanoma otomatis dari gambar kulit, penting untuk memiliki data pelatihan yang beragam dan melakukan pengujian terpilah berdasarkan berbagai demografis.

*FAT * '20, 27-30 Januari 2020, Barcelona, Spanyol*

978-1-4503-6936-7 / 20/01.
https://doi.org/10.1145/3351095.3372829

Tabel 1: Pelajaran dari Arsip: ringkasan pendekatan dalam ilmu arsip dan perpustakaan untuk beberapa topik paling penting dalam pengumpulan data, dan cara penerapannya dalam setelan pembelajaran mesin.

Persetujuan	(1) Melembagakan program penjangkauan pengumpulan data untuk secara aktif mengumpulkan data yang kurang terwakili (2) Mengadopsi model crowdsourcing yang mengumpulkan tanggapan terbuka dari peserta dan memberi mereka pilihan untuk menunjukkan sensitivitas dan akses
Inklusivitas	(1) Lengkapi set data dengan "Pernyataan Misi" yang menandakan komitmen terhadap konsep / topik / grup yang dinyatakan (2) set data "Terbuka" untuk mempromosikan pengumpulan berkelanjutan setelah pernyataan misi
Kekuasaan	(1) Membentuk konsorsium data di mana pusat data dengan berbagai ukuran dapat berbagi sumber daya dan beban biaya pengumpulan dan pengelolaan data
Transparansi	(1) Simpan catatan proses dari materi yang ditambahkan atau dipilih dari kumpulan data. (2) Mengadopsi sistem pengawasan data multi-lapisan dan multi-orang.
Etika & Privasi	(1) Mempromosikan pengumpulan data sebagai karir profesional penuh waktu. (2) Membentuk atau mengintegrasikan organisasi global / nasional yang ada dalam melembagakan kode etik / perilaku standar dan prosedur untuk meninjau pelanggaran

karakteristik untuk memastikan bahwa semua kelompok didiagnosis secara akurat. Pencarian kumpulan data yang besar dan representatif dapat menimbulkan pertanyaan tentang persetujuan yang diinformasikan. Keyes dkk. telah menunjukkan bahwa tolok ukur seperti yang berasal dari Institut Standar dan Teknologi Nasional di Amerika Serikat (NIST) terdiri dari data dari populasi rentan yang diambil tanpa persetujuan [38]. Pengujian terpilah juga memerlukan pengumpulan informasi yang berpotensi sensitif, dan mengkategorikan orang ke dalam berbagai kelompok berdasarkan informasi demografis (misalnya jenis kelamin, usia, ras, jenis kulit, etnis). Namun seringkali, tidak jelas bagaimana atau apakah orang harus dikategorikan sejak awal. Meskipun penting untuk mewakili orang dengan cara representasi yang mereka sukai (misalnya identitas gender), di lain waktu (seperti saat mendokumentasikan contoh-contoh diskriminasi),

Meskipun cara pengumpulan, anotasi, dan penggunaan data dalam ML memiliki konsekuensi yang luas, pengumpulan data belum diperiksa dengan teliti. Rangkuman Holstein dkk. Pada tahun 2019 tentang kebutuhan kritis untuk praktik yang adil di antara praktisi ML dalam industri mengidentifikasi kurangnya standar seluruh industri untuk "pengumpulan data yang sadar keadilan" sebagai area untuk perbaikan di seluruh bidang. Kurangnya proses sistematis untuk menghasilkan kumpulan data telah mendorong para peneliti untuk menyebutnya "barat liar" [28].

Baru-baru ini, fokus yang meningkat telah diberikan pada data, terutama dalam hal anotasi berbagai karakteristik demografis untuk pengujian terpilah, pengumpulan data yang representatif, dan penyediaan dokumentasi yang berkaitan dengan pengumpulan data dan proses anotasi [1, 21, 47]. Namun, langkah ini hanya mengatasi sebagian dari masalah. Masih ada pertanyaan terbuka tentang ketidakseimbangan kekuasaan, privasi, dan masalah etika lainnya. Sebagai peneliti menemukan lebih banyak masalah yang berkaitan dengan sistem ML, banyak yang mulai menyerukan pendekatan interdisipliner untuk memahami dan menangani masalah ini [58]. Demikian pula, kami menghimbau komunitas ML untuk mengambil pelajaran dari disiplin ilmu lain yang memiliki sejarah lebih lama dalam menangani masalah serupa. Secara khusus, kami fokus pada arsip, upaya manusia tertua untuk mengumpulkan data sosiokultural. Kami menguraikan kesejajaran arsip dengan upaya pengumpulan data di ML dan inspirasi dalam bahasa dan pendekatan kelembagaan untuk memecahkan masalah ini di ML. Karena arsip adalah lembaga yang terutama menangani dokumen dan foto, pelajaran ini paling baik diterapkan pada penggunaan subbidang data tidak terstruktur, seperti Natural Language Processing (NLP) dan Computer Vision (CV). Tentu saja, arsip hanyalah salah satu contoh dari bidang yang jauh yang dapat kita pelajari dari berbagai bidang. Dengan menunjukkan ketelitian yang diterapkan pada berbagai aspek pengumpulan data dan proses anotasi dalam arsip, sebuah industri sendiri, kami berharap dapat meyakinkan komunitas ML bahwa subbidang interdisipliner harus dibentuk dengan fokus pada pengumpulan data, berbagi, anotasi, pemantauan etika, dan proses pencatatan.

Karena disiplin ilmu yang terutama berkaitan dengan pengumpulan dokumentasi dan kategorisasi informasi, studi arsip telah menemukan banyak masalah yang terkait dengan persetujuan, privasi, keseimbangan kekuasaan, dan representasi, di antara masalah lain yang sekarang mulai dibahas komunitas ML. Sementara penelitian ML telah dilakukan dengan menggunakan berbagai tolok ukur tanpa mempertanyakan bias dalam set data, motif yang terkait dengan lembaga yang mengumpulkannya, dan bagaimana ciri-ciri ini membentuk tugas hilir, arsip memiliki

- pernyataan misi kelembagaan yang mendefinisikan konsep atau subkelompok untuk mengumpulkan data
- kurator penuh waktu yang bertanggung jawab untuk menimbang risiko dan manfaat mengumpulkan berbagai jenis data dan kerangka kerja teoretis untuk menilai data yang dikumpulkan
- kode etik / etika dan kerangka kerja profesional untuk menegakkannya
- bentuk standar dokumentasi yang mirip dengan apa yang diusulkan dalam Lembar Data untuk Kumpulan Data [21]

Selain itu, untuk mengatasi masalah representasi, inklusivitas dan ketidakseimbangan kekuasaan, ilmu kearsipan telah mendorong berbagai upaya kolektif seperti

- aktivisme berbasis komunitas untuk memastikan berbagai budaya terwakili dengan cara yang mereka inginkan (misalnya Mukurtu [2])
- konsorsium data untuk berbagi data antar lembaga untuk mengurangi biaya tenaga kerja dan infrastruktur.

Kami menyusun temuan kami tentang strategi pengarsipan menjadi 5 topik utama yang menjadi perhatian dalam komunitas ML yang adil: persetujuan, inklusivitas, kekuasaan, transparansi, dan etika & privasi. Tabel 1 merangkum pendekatan topik ini dalam studi arsip dan bagaimana mereka bisa melakukannya

2 mukurtu.org

diterapkan ke ML. Hasil kami menunjukkan bahwa arsip memiliki struktur kelembagaan dan prosedural yang mengatur pengumpulan, anotasi, dan penyimpanan data yang dapat diambil oleh ML.

Makalah lainnya disusun sebagai berikut. Bagian 2 memberikan gambaran umum tentang arsip dan relevansinya dengan ML. Bagian 3 membahas perbedaan tingkat pengawasan dalam ML dan pengumpulan data arsip. Bagian 4 membahas bagaimana pengumpulan data bisa lebih “intervensionis”. Bagian 5 menyajikan pendekatan arsip terhadap persetujuan, kekuasaan, inklusivitas, transparansi dan etika & privasi dan pelajaran yang dapat kita ambil darinya. Bagian 6 menjelaskan bagaimana kita dapat menerapkan pendekatan ini di tingkat masyarakat dan individu. Bagian 7 menyajikan studi kasus pengumpulan data untuk menggambarkan bagaimana konsep-konsep ini dapat diterapkan dalam praktik. Bagian 8 membahas batasan paralel dan aplikasi di ML. Bagian 9 diakhiri dengan pertanyaan dan tantangan terbuka.

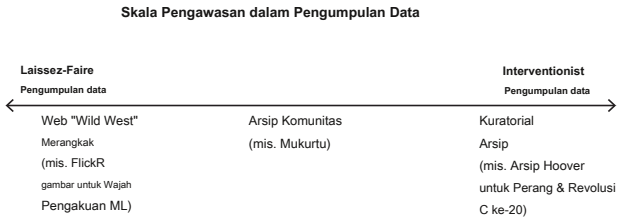
2 APA ITU ARSIP?

Arsip adalah koleksi bahan, sejarah dan terkini, disimpan secara sistematis untuk tujuan akademis, ilmiah, warisan, dan warisan. Sebagai bentuk pencatatan manusia kolektif berskala besar, arsip telah ada selama ribuan tahun, jauh sebelum materi digital. Arsip paling awal telah dilembagakan oleh negara dengan tujuan mengatur publik. Society of American Archivists (SAA) mendefinisikan arsip sebagai: Sebuah organisasi yang mengumpulkan catatan individu, keluarga, atau organisasi lain. [3]

Arsip mungkin bersifat institusional (mis. Arsip Perserikatan Bangsa-Bangsa [4]), pemerintah (mis. Administrasi Arsip dan Arsip Nasional [5]), dasar (mis. Rockefeller Archive Center [6]), berorientasi penelitian (mis. Houston Asian American Archive [7]), di antara memiliki tujuan lain. Banyak arsip modern memiliki komponen digital. Dalam semua kasus, arsip berbagi tujuan pengumpulan materi manusia sebagai catatan untuk dilihat untuk penggunaan di masa mendatang.

Melalui uji coba, tren, dan perdebatan selama bertahun-tahun, studi arsip memiliki literatur yang canggih tentang isu-isu yang menjadi perhatian dalam koleksi materi sosiokultural. Inisiatif keadilan baru-baru ini dalam prosedur dan bahasa gema komunitas ML telah dikembangkan dan digunakan dalam komunitas arsip dan perpustakaan. Beberapa di antaranya: pedoman tentang cara memberi label pada data [8] [57]; pengumpulan dan aksesibilitas informasi pribadi [13, 43]; berbagi dataset di seluruh platform [29, 61]; refleksi kritis tentang keragaman dan inklusivitas [24, 34]; teori penilaian dan seleksi [25]. Peneliti arsiparis telah mengembangkan berbagai sekolah pengumpulan data; TR Schellenberg, F. Gerald Ham, Terry Cook, dan Hans Booms berteori pendekatan yang berbeda untuk menilai dokumen [7]. Meskipun komponen digital pengarsipan masih baru, komunitas yang peka data di ML dapat memanfaatkan diskusi historis yang membahas pertanyaan mendasar tentang penggunaan dan penggalian informasi manusia.

3 archivists.org/glossary/terms/a/archives
4 search.archives.un.org
5 archives.gov
6 rockarch.org
7 haaa.rice.edu
8 github.com/saa-ts-dacs/dacs



Gambar 2: Contoh kategori praktik pengumpulan data pada skala supervisi.

3 PERBEDAAN ANTARA ARSIP DAN DATASET ML

Terlepas dari tujuan umum untuk mengumpulkan data atau informasi, arsip dan kumpulan data ML berbeda dalam beberapa dimensi. Mengidentifikasi perbedaan ini mendorong peneliti dan praktisi ML untuk melihat kemungkinan keragaman praktik pengumpulan data dan membekali mereka dengan kosakata untuk mengomunikasikan strategi pengumpulan.

Salah satu area di mana praktik pengumpulan data ML saat ini berbeda dengan yang ada di arsip kuratorial adalah tingkat intervensi dan pengawasan. Dalam praktiknya, pengumpulan data di subbidang ML yang signifikan dilakukan tanpa mengikuti prosedur atau serangkaian pedoman yang ketat. Sementara beberapa subbidang di ML memiliki pendekatan terperinci untuk pengumpulan data, subbidang lain seperti NLP dan CV menekankan ukuran dan efisiensi. Pendekatan ini sering mendorong pengumpulan data menjadi sembarangan [28]. Mengambil data dalam jumlah besar tanpa mengkritik asal, motivasi, platform, dan potensi dampaknya akan menghasilkan pengumpulan data yang diawasi secara minimal. Kami menunjukkan kemungkinan spektrum pengawasan pusat dalam strategi pengumpulan data dan arsip contoh di sepanjang sumbu pada Gambar. 2. Tidak ada satu titik pun pada spektrum yang benar-benar disukai dalam semua kasus tetapi tindakan ini berguna untuk mengetahui ukuran ini.

Arsip kuratorial terletak di sisi ekstrem lain dari spektrum intervensi. Raison d'être arsip ditentukan oleh Pernyataan Misi arsip (dibahas secara rinci di bagian 5.1), seringkali merupakan penggambaran yang sempit. Dalam arsip selektif, arsiparis dilatih untuk mengevaluasi dan *menyaring* sumber yang dianggap tidak relevan atau kurang berharga. Bagian 5.4 menunjukkan ada beberapa lapisan intervensi untuk menentukan apakah dokumen atau sumber tertentu layak ditambahkan ke koleksi. Mereka yang dianggap tidak memenuhi kriteria ini akan dihapus dari koleksi. Beberapa arsip juga menargetkan koleksi dari kelompok minoritas untuk mendiversifikasi koleksinya seperti yang didiskusikan di bagian 5.1. Kami menyebut metode “wild west” “laissez-faire” dan metode kuratorial sebagai pendekatan pengumpulan data yang lebih “intervensionis”.

Pengumpulan data ML dan arsip memiliki motivasi dan tujuan yang berbeda. Banyak set data ML memiliki tujuan akhir untuk memenuhi atau mengalahkan ukuran akurasi pada tugas yang ditentukan dari pelatihan model besar [28]. Kumpulan kuratorial bertujuan untuk melestarikan warisan dan ingatan, mendidik dan menginformasikan, dan secara historis cenderung peduli dengan keaslian, privasi, inklusivitas, dan kelangkaan sumber [12]. Namun masalah keaslian, privasi, dan inklusivitas ini juga harus menginformasikan penelitian ML dan menentukan tingkat pengawasan proyek dalam pengumpulan data dapat membantu menentukan tujuannya.

4 KEBUTUHAN UNTUK INTERVENTIONIS KOLEKSI

Meskipun mungkin ada pro dan kontra untuk pendekatan laissez-faire dan intervensi untuk pengumpulan data, kumpulan data yang disusun tanpa tingkat intervensi yang memadai akan mereplikasi bias yang timbul dari berbagai tingkat penyaringan. Bahkan sebelum pengumpulan, data tunduk pada dua tingkat bias - historis dan representasi [60]. Untuk meminimalkan bias ini, strategi pengumpulan data harus melakukan intervensi sebelum menerapkan pengambilan sampel, pembobotan, dan teknik penyeimbangan lainnya.

Bias historis mewakili ketidakadilan struktural dan empiris yang melekat pada masyarakat yang tercermin dalam data, seperti kurangnya historis presiden perempuan di banyak negara dan kurangnya representasi ras minoritas dalam kepemimpinan bisnis. Jika diambil secara grosir, kumpulan data ML yang besar dan tidak pandang bulu menghasilkan turunan dan hasil yang mencerminkan bias ini. Misalnya, [19] menunjukkan embeddings kata terlatih yang mereplikasi stereotip Asia dalam data bahasa dari Google Buku historis dan Corpus of Historical American English (COHA).

Bias representasi berasal dari divergensi antara distribusi sebenarnya dan ruang input digital. Hal ini dapat diakibatkan oleh akses yang tidak merata ke perangkat digital atau kendala sosial budaya yang mencegah digitalisasi atau pelestarian. Misalnya, wanita di Uni Emirat Arab secara sosial distigmatisasi agar tidak memotret wajah mereka, sehingga ketersediaan jenis data ini menjadi lebih rendah daripada distribusi sebenarnya. Beberapa materi mungkin sengaja dihancurkan untuk tujuan politik. Pada akhir Perang Dunia II, petugas Nazi Gestapo membakar catatan organisasi dan prosedur mereka untuk menghindari penuntutan [6]. Contoh-contoh ini membiaskan ketersediaan jenis data tertentu.

Set data "alami", seperti yang dirayapi dari internet, harus memiliki lapisan intervensionis untuk mengatasi ketidakadilan ini paling baik dan setidaknya digunakan dengan hati-hati. Banyak proyek ML dilatih pada kumpulan data berdasarkan materi yang ditemukan di internet atau sudah dalam format digital. Sistem pengenalan wajah komersial telah menggunakan FlickR sebagai sumber gambar wajah alami manusia [59]. Sumber umum bahasa alami manusia di NLP mencakup materi yang bersumber dari banyak orang seperti teks Twitter atau data dari situs platform publik seperti Yelp, Wikipedia, IMDB, Stackoverflow, dan Reddit. Diambil tanpa intervensi, kumpulan data ini mengalami sumber bias di atas. Pertama, materi yang ditemukan di internet mencerminkan komposisi demografis tertentu.

Namun menggunakan bahan dari sumber yang telah melalui pengawasan manusia tidak selalu "aman". Misalnya, kumpulan data umum lainnya di NLP adalah teks dari situs berita online (mis. Jurnal Wall Street, BBC, CNN, Reuters). Meskipun data berita yang dipublikasikan melalui proses penyaringan yang diawasi oleh dewan editorial, hal ini tidak menghalangi mereka untuk memiliki bias politik dan topik. Fox News cenderung menulis untuk kepentingan konservatif politik AS dan Wall Street Journal cenderung menghasilkan karya tentang topik bisnis dan perdagangan yang relevan, umumnya mempromosikan perdagangan bebas dan kebijakan ekonomi liberal lainnya. Christian Science Monitor memiliki rasa religius, melayani audiens yang senama. Integral dengan komposisi set data berita adalah satu-satunya tujuan penyedia berita komersial

Tabel 2: Paralel dalam model pengarsipan dan pengumpulan data.

Arsip	Pengumpulan Data ML
Kebijakan Koleksi & Pernyataan Misi	Tujuan Model, Kebutuhan yang berpusat pada pengguna [51]
Konsorsium Arsip	Data Trust [32]
Arsip Partisipatif & Komunitas	Sumber daya kerumunan [5]
Pedoman & Catatan Penilaian Kode	Kode Etik AI [35] Lembar Data
Etik & Perilaku	untuk Set Data [21]

untuk menjaga pembaca dan pangsa pasar yang sehat. Mereka tidak menghasilkan materi yang menyadari dampaknya pada model ML - hanya peneliti ML yang menggunakan materi ini yang akan memiliki kepemilikan atasnya. Investigasi kritis dari motivasi dan tujuan data merupakan komponen penting dari pengumpulan data intervensionis.

Seperti dibahas di bagian 5, ada beberapa strategi untuk menarik dalam pengumpulan data untuk mengurangi tingkat bias ini. Kami mendorong para peneliti dan institusi untuk lebih sadar dan aktif dalam pengumpulan data.

5 PELAJARAN DARI ARSIP

Kami menyusun isu-isu dalam transparansi akuntabilitas keadilan dan etika menjadi 5 abstraksi utama dan menunjukkan bagaimana arsip telah mendekati mereka melalui cara kelembagaan dan intervensionis. Beberapa model ini memiliki kesamaan dalam inisiatif terbaru di komunitas ML seperti yang ditunjukkan pada Tabel 2, memungkinkan kami untuk mempertimbangkan keberhasilan dan kegagalan dari setiap pendekatan dalam konteks ML.

5.1 Inklusivitas: Pernyataan Misi & Kebijakan Koleksi

Inklusivitas telah menjadi masalah dalam ML karena praktik pengumpulan data tidak didorong oleh agenda yang memprioritaskan representasi atau keragaman yang adil, melainkan oleh tugas atau kenyamanan. Dalam ML, banyak set data dikumpulkan untuk tugas Artificial Intelligence (AI) tertentu. Misalnya, di NLP, kumpulan data PennTreebank (berdasarkan Wall Street Journal) telah berfungsi selama bertahun-tahun sebagai standar untuk tag Part-of-Speech [45]. Lainnya termasuk OntoNotes untuk resolusi coreference dan pengenalan entitas bernama (NER) dan korpus paralel EuroParl untuk terjemahan mesin (MT) [30, 40]. Kadang-kadang, peneliti dan praktisi membangun di atas kumpulan data yang ada yang dimotivasi oleh ketersediaan dan kenyamanan. Berdasarkan judul makalah yang diterima di Asosiasi Linguistik Komputasi (NAACL) 2019 Bab Amerika Utara, [9] Pendekatan ini menghasilkan kumpulan data berbasis sumber, di mana metode pengumpulan atau pertanyaan ditentukan oleh ketersediaan kumpulan data.

Menetapkan agenda pengumpulan data dengan ketersediaan digital menghasilkan data yang bias seperti yang dibahas di bagian 4 dan mereplikasi bias ini dalam model. Dibiarkan tanpa pengelolaan aktif komposisi data, metode ini dapat menyebabkan terbatasnya cakupan demografis (misalnya jenis pengguna apa yang menggunakan Reddit?) Dan membuat model yang dihasilkan menjadi sangat bergantung pada sumber tertentu (misalnya model ini dilatih pada data Reddit sehingga mencerminkan pengguna Reddit). Peneliti telah meliput konsekuensi dari ML yang dilatih pada set data yang tidak memiliki keragaman dan menyerukan regulasi yang lebih baik [8, 36].

[9] naacl2019.org/program/accepted

Arsip menghadapi kritik serupa terhadap eksklusivitas. Arsip tradisional lebih menitikberatkan pada dokumen negara dan pemerintahan, dengan tujuan untuk melestarikan dokumen elit pemerintahan dan sosial. Baru setelah kebangkitan sejarah sosial tahun 1960-an, arsip mulai merekam kehidupan non-elit dengan sungguh-sungguh [27]. Arsip membahas inklusivitas dengan lebih berhati-hati dan sadar dalam menentukan tujuan pengumpulan data. Salah satu arsip strategi yang dipatuhi adalah memiliki Pernyataan Misi. Daripada memulai dengan kumpulan data berdasarkan ketersediaan, pengumpulan data dalam arsip dimulai dengan pernyataan komitmen untuk mengumpulkan peninggalan budaya dari konsep, topik, atau kelompok demografis tertentu. Pernyataan atau kebijakan pengumpulan ini dapat menargetkan spesialisasi seperti minoritas gender di New York atau dokumen tentang sejarah Amerika Barat. Banyak yang secara aktif mengumumkan komitmen terhadap koleksi minoritas. Semua arsip memiliki Pernyataan Misi panduan. Beberapa contohnya di bawah ini:

> RadcliffeCollege / Schlesinger Library Schlesinger
> Misi perpustakaan adalah mendokumentasikan kehidupan perempuan dari masa lalu dan masa kini untuk masa depan. Kepemilikannya menerangi sejumlah besar individu, keluarga, organisasi, peristiwa, dan tren dan berisi kekayaan sumber daya untuk mempelajari sejarah sosial, politik, ekonomi, dan budaya. Koleksi perpustakaan sangat kaya di bidang hak-hak dan feminisme perempuan, kesehatan dan seksualitas, reformasi dan aktivisme sosial, pekerjaan dan kehidupan keluarga, sejarah kuliner, serta pendidikan dan profesi. [10]

> Memori Kerajaan Pedalaman Inland Empire Memories adalah aliansi perpustakaan, arsip, dan organisasi warisan budaya yang didedikasikan untuk mengidentifikasi, melestarikan, menafsirkan, dan berbagi budaya yang kaya dari berbagai komunitas di Riverside dan San BernardinoCounties, wilayah geografis yang juga dikenal sebagai California Selatan Pedalaman. Inisiatif ini berupaya untuk meningkatkan akses ke catatan utama individu dan organisasi yang pekerjaannya secara fundamental membentuk pengalaman hidup orang-orang di Inland Southern California. Penekanan khusus akan diberikan pada bahan-bahan yang mendokumentasikan kehidupan masyarakat dan kelompok yang kurang terwakili dalam catatan sejarah.
>
> [11]

Pengumpulan data yang didorong oleh konsep lebih mahal dan memakan waktu. Ini membutuhkan keahlian domain (mis. Kelompok minoritas gender apa yang harus kita pertimbangkan?) Dan eksplorasi (mis. Organisasi lokal mana yang harus kita hubungi?). Tetapi Pernyataan Misi publik memaksa para peneliti untuk memperhitungkan komposisi data mereka dengan memandu proses pengumpulan data. Pencarian diperluas dalam berbagai sumber (misalnya, di mana saya bisa mendapatkan gambar upacara pernikahan lintas budaya?) Tetapi juga filter (misalnya, apakah dokumen ini relevan dengan Pernyataan Misi proyek?). Komunitas arsip membuka dialog terbuka tentang praktik-praktik yang baik untuk menyusun Pernyataan Misi [39].

Mengumumkan Pernyataan Misi publik juga memungkinkan kumpulan data terbuka untuk kontribusi berkelanjutan. Peneliti dapat mengupdate dataset berdasarkan perubahan norma sosial budaya. Menggabungkan kumpulan data dengan Pernyataan misi dapat menginformasikan komunitas riset tentang kepemilikan mereka dan amandemen data di masa mendatang. Misalnya, dataset gambar dengan pernyataan komisi seperti *"Mengumpulkan gambar wanita dalam berbagai pekerjaan"* dapat mengundang komunitas untuk terus menyumbangkan data yang memenuhi tujuan pernyataan tersebut.

5.2 Persetujuan: Arsip Komunitas & Partisipatif

Dalam ML, crowdsourcing telah muncul sebagai pendekatan pokok dalam mengumpulkan label manusia untuk kumpulan data yang bertujuan untuk mengurangi biaya dan mempercepat pengumpulan data dengan melakukan outsourcing ke peserta manusia. Project Re- spect crowdsources istilah sentimen positif terkait dengan komunitas LGBTQ + untuk menyeimbangkan asosiasi linguistik negatif yang terkait dengan gender minoritas online. [12] Namun, ada beberapa contoh di mana crowdsourcing ML mengandalkan masukan terbuka dari masyarakat. Sebagian besar mekanisme crowdsourcing menyediakan satu set label tetap bagi peserta untuk dipilih, membatasi peserta untuk berkontribusi pada pilihan terbatas yang ditentukan oleh agenda peneliti. Beberapa proyek crowdsourcing menyiratkan hubungan permusuhan antara peneliti dan peserta, mengusulkan berbagai pengaturan eksperimental yang menampilkan kompetisi dan model insentif. [13]

Metode ini terus menghadapi kritik karena kurang keragaman dan memaksakan label pada individu. Peneliti ML tanpa pengetahuan domain yang memadai dari kelompok minoritas sering salah mengartikan data, memaksakan label yang tidak diinginkan atau bahkan merugikan ke dalam kelompok. Salah satu contoh paling menonjol adalah gender. Keyes membahas bahaya sosial ekonomi dari pelabelan gender pada transgender [37]. Hamidi dkk. menunjukkan keberatan yang luar biasa terhadap sistem Pengakuan Gender Otomatis di antara komunitas non-biner dan non-gender yang sesuai [26]. Seringkali, label-label ini berubah dari waktu ke waktu, menunjukkan tingkat subjektivitas sosial mereka. Misalnya, kategori ras yang direstui negara telah berkembang di Amerika Serikat. Dalam sensus AS tahun 1860, seseorang dapat memilih untuk menandai di antara opsi, "W" (Putih), "B" (Hitam), atau "M" (Mulattos) (AS Biro Sensus 1860). Pada tahun 1890, opsi diperluas menjadi "Putih", "Hitam", "Mulatto", "Quadroon", "Oktoroon", "China", "Jepang", atau "India". (Biro Sensus AS 1890).

Arsip komunitas, juga disebut arsip suku atau arsip partisipatif, adalah proyek pengumpulan data atau dokumen milik kelompok yang diwakili. Proyek seperti historypin [14] menyediakan platform bagi komunitas lokal untuk menentukan dan menyumbangkan koleksi warisan dan budaya mereka sendiri. Tujuannya adalah untuk memperluas saluran input pengguna dalam pengumpulan data. Arsip komunitas dimotivasi oleh kebutuhan untuk mewakili suara-suara “non-elit, akar rumput, yang terpinggirkan” [17]. Di dunia telepon, arsip komunitas telah mendokumentasikan kelompok-kelompok minoritas sejak tahun 1970-an mulai dari arsip LGBT (Hall -Carpenter Archives 1980s) ke Black Cultural Archive yang didirikan di

Inggris pada tahun 1981. Ribuan arsip pengumpulan-sendiri lainnya ada saat ini yang mencakup berbagai kelompok agama, bahasa, kelas, jenis kelamin, etnis, generasi, budaya, dan regional yang dibantu oleh platform online. Di Inggris saja, lebih dari satu juta orang dilaporkan terlibat

[10] radcliffe.harvard.edu/schlesinger-history-and-holdings
[11] inlandempirememories.org/mission
[12] projectrespect.withgoogle.com
[13] crowdml.cc/nips2016
[14] historypin.org/en

dalam pengarsipan komunitas [27]. Tabel 3 mencantumkan contoh arsip komunitas. [15]

Arsip komunitas berfungsi sebagai contoh bagaimana dataset dapat dibuka untuk masukan publik, mendemokratisasi proses pengumpulan, dan memberikan lembaga kepada kelompok minoritas untuk mewakili diri mereka sendiri. Untuk tujuan pengarsipan sejarah, misi untuk membangun warisan lokal yang inklusif dan melestarikan akun terlengkap mungkin mendasari inisiatif ini. Misalnya, arsip wanita seperti Feminist Archive (1978-), kumpulan diari, surat pribadi, foto, di antara ephemera lainnya, berkontribusi pada sejarah nasional yang lebih lengkap. Ribuan arsip komunitas semacam itu menambah keragaman pada catatan sejarah.

Sementara banyak dari inisiatif yang dimulai dari akar rumput, yayasan dan lembaga telah secara aktif mendanai dan mempromosikan pengarsipan di pinggiran. Misalnya, ada yang bertujuan untuk meningkatkan partisipasi dengan mendistribusikan “kit” peralatan ke daerah pedesaan. Kit ini mencakup peralatan untuk mendigitalkan data audio dan video yang berisiko dan peralatan untuk menyelamatkan bahan dari penyimpanan usang [50]. Program semacam itu secara aktif mengirimkan sumber daya ke kelompok lokal dan minoritas untuk mengumpulkan kontribusi mereka. Kami belum melihat inisiatif serupa untuk mengumpulkan lebih banyak jenis data di ML.

Model desentralisasi ini juga memungkinkan kelompok minoritas untuk menyetujui dan menentukan kategorisasi mereka sendiri. Beberapa budaya menuntut sistem representasi non-Barat. Mukurtu adalah contoh sistem manajemen konten yang dibuat untuk memungkinkan komunitas adat menyimpan materi mereka sendiri "menggunakan protokol mereka sendiri." Didanai oleh National Endowment for the Humanities dan Institute of Museum and Library Services, Mukurtu menyediakan platform bagi individu untuk mengunggah data mereka, menandai konten sensitif, dan memberi label preferensi mereka untuk akses, penggunaan, dan sirkulasi semua melalui protokol yang dikontrol secara ketat . Proyek-proyek ini memberikan lembaga kepada penduduk asli untuk menggunakan kosa kata mereka sendiri sebagai label dan pilihan gambar mereka untuk diintegrasikan ke dalam database Mukurtu. Gambar 3 adalah contoh gambar yang diunggah dan diberi label oleh penduduk setempat [9].

Meskipun set data ML cenderung lebih besar dan homogen, contoh ini dapat berfungsi sebagai template awal untuk prosedur pengumpulan data ML yang berpusat pada komunitas di masa mendatang. Misalnya, proyek crowd-sourcing ML dapat mengatur struktur analog di sekitar pengumpulan data partisipatif yang secara lebih aktif membekali peserta dengan pilihan untuk akses, sirkulasi, dan sensitivitas. Seperti Mukurtu, dataset ML yang mengumpulkan konten budaya internasional misalnya, bisa didesain sedemikian rupa sehingga partisipan menandai sensitivitas dan memberi masukan terbuka.

Tentu saja, desentralisasi pengumpulan data dapat memperluas masukan publik tetapi juga menimbulkan tantangan. Arsip komunitas telah menghadapi penolakan bahwa kualitas data berkurang karena kumpulan kontributor semakin besar [42]. Sudut pandang yang terpisah ini tentang berbagai tingkat partisipasi publik dalam pengumpulan data memperingatkan para peneliti ML untuk benar-benar memeriksa paradigma pengumpulan data mereka sendiri. Per proyek, peneliti ML harus bertanya, seberapa besar pengawasan, keahlian domain, dan spesialisasi yang dibutuhkan dalam mengumpulkan data untuk proyek cakupan yang ada. Misalnya, peneliti NLP yang melatih kata embeddings pada dialek regional mungkin ingin bekerja sama dengan antropolog atau

[15] gerberhart.org; lesbianherstoryarchives.org; saha.org.za; saada.org; wcml.org.uk; wisearchive.co.uk

Tabel 3: Contoh Arsip Komunitas

Seksualitas & Gender	Politik	Kelas sosial
Perpustakaan Gerber / Hart dan Arsip	Afrika Selatan Arsip Sejarah	Kelas pekerja Perpustakaan Gerakan
Herstory Lesbian Arsip	Asia Selatan American Digital Arsip	WISEArchive

Gambar 3: Gambar tarian tradisional Catawba yang diambil pada tahun 1993 di reservasi Catawba Indian. © Proyek Pelestarian Budaya Catawba

ahli domain lainnya untuk menjangkau subkelompok yang benar dan mengakomodasi perbedaan budaya.

5.3 Kekuasaan: Konsorsium Data

Menerapkan sistem pengumpulan data etis menuntut waktu, keahlian, dan sumber daya. Melakukan pengujian terpilah membutuhkan biaya tenaga kerja anotator untuk label demografis tambahan dan menangani data sensitif dengan hati-hati membutuhkan sumber daya, keahlian dan infrastruktur untuk menjaga privasi. Semua ini membutuhkan institusi yang merugikan dengan sumber daya yang rendah. Seperti dilaporkan dalam [41], saham perusahaan rintisan yang lebih kecil jatuh setelah pengumuman Peraturan Perlindungan Data Umum Eropa (GDPR) karena perusahaan teknologi yang lebih besar dapat memanfaatkan lebih banyak sumber daya untuk memastikan kepatuhan GDPR daripada lembaga yang lebih kecil.

Untuk meningkatkan paritas dalam kepemilikan data, arsip dan perpustakaan telah mengembangkan model konsorsium. Pada awal abad kedua puluh, kelompok arsip dan perpustakaan mendirikan kerangka kerja dan layanan kelembagaan untuk berbagi sumber daya dan secara kolektif menyimpan dan mendistribusikan kepemilikan yang disebut jaringan perpustakaan, koperasi, dan konsorsium [29]. Contoh-contoh termasuk OCLC, LYRASIS, AMIGOS Library Services, OhioLINK, dan MELSA [29]. Pada September 2003, International Coali- tion of Library Consortia (ICLC) telah mengutip lebih dari 170 konsorsium sebagai anggota yang lebih dari 100 di antaranya berbasis di AS [14].

Konsorsium memiliki beberapa keuntungan timbal balik bagi kelompok yang berpartisipasi. Keuntungan utama adalah kemampuan untuk mendapatkan skala ekonomi. Grup perpustakaan dapat melakukan pembelian mahal seperti langganan jurnal akademik, mengumpulkan sumber daya untuk proyek skala besar seperti HathiTrust dan DPLA, dan mengurangi biaya overhead teknologi.

Juga, dengan mengkomunikasikan kepemilikan dengan lembaga lain, anggota konsorsium dapat secara kolektif mengurangi koleksi yang berlebihan, alih-alih berfokus pada peningkatan ukuran koleksi unik. Institusi yang lebih kecil juga bisa mendapatkan keuntungan dari bergabung dengan konsorsium bergengsi dengan meningkatkan visibilitas ke kepemilikan unik bersama mereka. Banyak perpustakaan berpartisipasi dalam beberapa konsorsium: Universitas Negeri Carolina Utara adalah anggota dari sembilan konsorsium. Konsorsium yang efektif menguntungkan lembaga partisipan yang lebih kecil dan lebih besar dengan memperbesar ukuran koleksi konsorsium dan membuat proyek yang tidak mungkin menjadi mungkin [29].

Tetapi sejarah telah menunjukkan bahwa konsorsium bukannya tanpa kekurangan. Pada 1990-an dan 2000-an, konsorsium mendapat kecaman karena menciptakan birokrasi, komite yang tidak perlu, penundaan, dan bentuk-bentuk baru ketidakseimbangan kekuasaan di antara anggota konsorsium. Konsorsium didanai oleh kontribusi keanggotaan di mana para anggotanya dapat memiliki berbagai tingkat kemampuan keuangan, yang menentukan potensi ketidakadilan kekuasaan. Tantangan tambahan konsorsium di ranah data ML adalah hubungan rumit antara laba dan data. Banyak organisasi teknologi besar memiliki kumpulan data eksklusif yang mungkin tidak mereka bagikan dalam pengaturan konsorsium.

Komunitas ML secara aktif mendiskusikan pengaturan konsorsium untuk berbagi data. Di Inggris Raya, inisiatif kepercayaan data dari Open Data Institute (ODI) sedang dalam pengembangan tetapi masih pada tahap mendefinisikan fungsi dan kapasitasnya. [16] Model awal ini dapat belajar dari uji coba dan kesalahan konsorsium perpustakaan.

5.4 Transparansi: Catatan Penilaian & Pengumpulan Data berbasis Komite

Komunitas keadilan ML telah mengusulkan berbagai langkah untuk mengatasi kurangnya transparansi dalam pengumpulan data dan arsitektur model ML [1, 21, 47]. Proposal ini mendorong komunikasi yang jelas tentang bahan dan prosedur yang membentuk proyek ML dengan publik. Misalnya, di Lembar Data untuk Kumpulan Data, penulis menghitung pertanyaan yang meminta peneliti untuk membahas bagaimana kumpulan data yang diberikan dikumpulkan [21]. Transparansi dan akuntabilitas membentuk prinsip sentral etika kearsipan [2, 3, 56]. Untuk menegakkan prinsip-prinsip ini, arsip mematuhi standar pencatatan yang ketat, tidak hanya meningkatkan transparansi tetapi juga operasi yang efektif dengan lembaga lain.

Seperti lembar data, arsip telah mengembangkan standar terperinci untuk deskripsi data. Arsip menggunakan tiga kategori standar untuk mengkomunikasikan kepemilikan yang konsisten di seluruh lembaga: 1) Standar konten data yang menentukan konten, urutan, dan sintaks data (ISADG, DACS, RAD), 2) standar struktur data, menentukan organisasi data (EAD, EAC-CPF), dan 3) standar nilai data, menentukan istilah yang digunakan untuk menggambarkan data (LCSH, AAT, NACO) [7]. Bagian dari tugas arsiparis adalah menyimpan catatan yang sesuai dengan standar ini.

Namun selain merinci isi datanya, arsip juga mencatat *proses* pengumpulan data. Arsip berhati-hati karena semua konten dan catatan arsip pada akhirnya akan melayani generasi mendatang. Dengan premis ini, arsiparis menyimpan catatan keputusan dan evaluasi aliran appraisal. Dalam penilaian yang teliti, proses tersebut melewati banyak lapisan pengawasan oleh arsiparis, kurator, pembuat arsip, dan manajer arsip. Tabel 4 menunjukkan multi-level dan multi-person contoh proses penilaian. [17] Level-level yang sepadan dengan 1 dan 2 dalam contoh ini berbasis komite dan level 3 dan 4 didelegasikan kepada kurator dan pemroses profesional.

Tabel 4: Contoh Alur Penilaian (Arsip Hoover)

1. Pernyataan Misi
- Tingkat tertinggi perumusan agenda yang menentukan topik / konsep yang menjadi perhatian.

2. Kebijakan Pengembangan Koleksi
- Kebijakan yang lebih spesifik diambil dari Pernyataan Misi tentang apa yang dikumpulkan, apa yang tidak, dan di mana dan bagaimana mencari sumber.

3. Penilaian
- Evaluasi berdasarkan kriteria apakah pilihan sumber tertentu layak dikumpulkan.
- Menanyakan apakah koleksi ini sesuai dengan garis besar pernyataan misi
- Mengevaluasi kelangkaan sumbernya, keaslian asalnya, dan nilainya untuk generasi mendatang.

4. Pengolahan / Pengindeksan (Micro-Appraisal)
- Memproses sumber secara individual atau pada tingkat folder / dokumen, termasuk mengindeksnya dan memperbarui alat bantu pencarian.
- Sumber dapat dibuang karena masalah privasi atau karena tidak relevan.

Berbagai tingkat tinjauan dan penyimpanan catatan ini tidak pernah terdengar dalam pengumpulan data ML. Sementara memperkenalkan langkah-langkah ini dalam prosedur pengumpulan data menambah biaya dan keterlambatan dalam pengembangan, contoh dari arsip ini berfungsi sebagai model untuk strategi masa depan secara adil.
ML.

5.5 Etika & Privasi: Kode Etik dan Perilaku

Para peneliti telah mengusulkan metode dan organisasi untuk mengatur standar etika dalam AI, menyoroti kebutuhan privasi untuk monitor dan akuisisi data yang etis [10, 15]. Di ranah teknologi konsumen, pemberlakuan GDPR pada tahun 2018 di Uni Eropa (UE) menandai titik balik dalam sejarah perlindungan data konsumen dengan menerapkan penalti dan sanksi ketidakpatuhan pada bisnis. Tidak ada bentuk regulasi yang serupa di Amerika Serikat. Secara global, ISO-IEC JTC 1 yang didirikan pada tahun 1987 dan Institute of Electrical and Electronics Engineers (IEEE) Global inisiatif on Ethics of Autonomous & Intelligent Systems adalah dua standar utama untuk penelitian dan pengembangan AI. Namun langkah-langkah ini tidak membahas etika pengumpulan data apalagi menyediakan mekanisme penegakan hukum. Standar, Terlepas dari pentingnya, menjadi lebih sulit untuk diterapkan karena tujuan bergerak semakin jauh dari transaksi pasar akhir dan praktik etis dalam pengumpulan data sering diabaikan karena jaraknya dari produk akhir. [10].

[16] theodi.org

[17] Wawancara dengan arsiparis Hoover dilakukan pada 4/10/2019 di Stanford, California, Amerika Serikat

Masalah etika yang terkait dengan pengumpulan data sosiokultural memiliki sejarah panjang dalam arsip. Penanganan informasi manusia membuat arsiparis menghadapi berbagai dilema etika: memilih dokumen mana yang akan dibuang atau disimpan, memberikan akses ke konten sensitif, dan berurusan dengan kekayaan intelektual hanyalah beberapa dari banyak [18, 46, 53, 55] . Beberapa lapisan kode yang tumpang tindih pada panduan perilaku profesional dan menegakkan keputusan tentang masalah ini. Organisasi payung atas arsip, perpustakaan, dan museum masing-masing memiliki kode etik dan perilaku masing-masing. Kode etik arsip melalui SAA mencantumkan nilai inti arsiparis untuk meningkatkan akses, memastikan akuntabilitas dan transparansi, melestarikan beragam bahan, memilih bahan secara bertanggung jawab, dan menyimpan catatan demi generasi mendatang [56]. [18] Kelompok internasional seperti Dewan Internasional tentang Arsip (ICA) dan Dewan Museum Internasional (ICOM) memelihara protokol etika yang diperbarui yang diterjemahkan ke dalam banyak bahasa untuk mempromosikan praktik global standar.

Penegakan etika dalam ML menghadapi tantangan karena kurangnya mekanisme insentif bagi peneliti dan praktisi. Meskipun arsip mungkin tidak menemukan strategi penegakan hukum yang sangat mudah, ada beberapa ciri ekosistem arsip yang memaksa arsiparis untuk mematuhi pedoman etika. Pertama, sebagian besar arsiparis adalah pengumpul data profesional penuh waktu. Di banyak organisasi, arsip bekerja dengan sistem keanggotaan di mana pelanggaran kode etik dapat mengakibatkan hilangnya keanggotaan profesional [3]. Banyak sub-organisasi pengarsipan dan pengumpul catatan memiliki panel atau komite etika yang mengevaluasi setiap dugaan pelanggaran kasus per kasus [3, 31]. Dalam ML, mempromosikan pekerjaan penuh waktu dalam pengumpulan data akan memperkenalkan cara meningkatkan insentif bagi pengumpul data untuk mematuhi standar etika. Ketika tugas utama pengumpul data adalah memilih dan mengevaluasi data di bawah kode etik dan keanggotaan profesional mereka bergantung pada tugas ini, kepatuhan mungkin lebih mudah untuk diterapkan. Karena pengumpulan data itu sendiri adalah pekerjaan tanpa akhir, kode etik dapat secara signifikan memandu pekerjaan pengumpul data seperti yang dilakukan oleh seorang arsiparis.

Kedua, membangun organisasi lintas-kelembagaan yang berada di atas pemberi kerja langsung dapat membantu memastikan bahwa prinsip-prinsip etika tahan terhadap motif yang digerakkan oleh keuntungan. Komitmen anggota individu terhadap kode etik memberdayakan pengumpul data untuk menahan tekanan dari majikan mereka untuk mengambil jalan pintas. Sementara banyak perusahaan sudah mulai mengukir Prinsip AI [19], dan organisasi seperti Kemitraan dalam AI telah dibentuk untuk menciptakan standar lintas kelembagaan [20], mereka hanya akan efektif jika ada mekanisme di mana lembaga dimintai pertanggungjawaban. Komunitas ML dapat belajar dari strategi penegakan arsip.

6 DUA TINGKAT TINDAKAN

Kami dapat mengambil tindakan di tingkat makro dan mikro untuk meningkatkan pengumpulan dan anotasi data ML. Di tingkat makro, sebagai komunitas, lembaga swasta, pembuat kebijakan, dan organisasi pemerintah, kita dapat:

(1) Menggabungkan dan mengembangkan konsorsium data
(2) Membentuk organisasi profesional yang bekerja oleh anggota-kapal untuk menegakkan kepatuhan terhadap pedoman etika (3) Mendukung arsip komunitas
(4) Mengembangkan subbidang yang didedikasikan untuk pengumpulan data dan proses notasi

Di tingkat mikro, sebagai peneliti individu, praktisi, dan administrator, kami dapat:

(1) Tentukan dan modifikasi Pernyataan Misi
(2) Mempekerjakan staf penuh waktu untuk pengumpulan data yang kinerjanya bagus juga terikat dengan standar profesional (3) Bekerja menuju kumpulan data publik
(4) Mengadopsi standar dokumentasi dan menyimpan dokumen yang ketat pemikiran
(5) Mengembangkan kebijakan pengembangan koleksi yang lebih substansial dengan keahlian domain dan nuansa sumber data
(6) Membuat keputusan komite berdasarkan data diskresioner

Kedua tingkat ini memperkuat dan mendukung satu sama lain. Misalnya, akan lebih efektif untuk mengelola prinsip-prinsip etika ketika organisasi pusat di luar proyek individu meminta pertanggungjawaban pengumpul data.

7 STUDI KASUS: GPT-2 DAN REDDIT

Untuk menggambarkan bagaimana pelajaran dari sejarah arsip dapat diterapkan pada pengumpulan data ML dalam praktiknya, kami mengambil contoh WebText dari model bahasa GPT-2 OpenAI sebagai studi kasus [54]. Kami memilih GPT-2 untuk studi kasus kami karena rilis yang tepat waktu, perhatian penulis terhadap implikasi etis dari merilis model bahasa besar [11], dan rilis model card mereka [11] yang mendokumentasikan kinerja GPT-2 dan kesesuaiannya kasus penggunaan yang memperingatkan pengguna tentang bias dalam data pelatihan. Kami mengeksplorasi motivasi untuk pendekatan khusus Web-Text dan memberikan saran untuk perbaikan sepanjang 5 dimensi yang kita diskusikan. MeskipunWebText tidak tersedia untuk umum, untuk tujuan studi kasus ini, kami anggap demikian.

GPT-2 dilatih di WebText, kumpulan 8 juta dokumen (~ 40G) dikumpulkan dengan menghapus tautan dari halaman Reddit dengan setidaknya 3 suara positif bersih. Itu *dinyatakan* motivasi untuk proses pengumpulan data ini adalah tujuan kinerja yang melatih model bahasa untuk ditransfer ke berbagai tugas NLP. Berbeda dengan kumpulan data NLP khusus domain, fokus WebText adalah mengumpulkan kumpulan data menjadi besar dan beragam, mencakup berbagai konteks, dengan penekanan pada kualitas data. Diasumsikan sebagian besar konten berbahasa Inggris karena situs tersebut beroperasi Anglophone [54].

Selain itu, masih ada beberapa lainnya *tidak dinyatakan* tetapi motivasi dapat disimpulkan untuk pendekatan WebText. Pertama, hemat biaya dan waktu karena data yang dikumpulkan bersifat publik dan online. Kedua, berpusat pada Reddit meningkatkan kemungkinan bahwa kumpulan data tersebut kontemporer dan berasal dari distribusi yang serupa dengan data validasi, yang sering juga merupakan kumpulan data yang di-web-scrap. Ketiga, Reddit adalah platform yang cukup dikenal luas di komunitas ML sehingga kemungkinan tidak memerlukan justifikasi ekstensif penggunaan atau penjelasan konten.

Dengan kata lain, kesimpulan kami adalah bahwa WebText tidak dioptimalkan untuk inklusivitas sosiokultural. Motivasi untuk merakit Web-Teks adalah untuk meningkatkan performa model ML yang diusulkan

[18] atsilirn.aiatsis.gov.au/protocols.php
[19] futureoflife.org/ai-principles
[20] partnershiponai.org

- karena itu fokus pada keragaman konteks dan mode linguistik - bukan untuk melatih model kelas industri - sehingga tidak memperhatikan keragaman budaya.

Akibatnya dari perspektif sosiokultural, komposisi WebText menyarankan satu kendala keras dan pembatasan yang lebih lembut. Kendala utama adalah bahwa semua konten seluruhnya dari dokumen yang ditemukan di internet dengan tautan aktif dan dapat diakses publik. Ini membatasi sumber untuk menjadi bahasa tertulis yang ditemukan secara online baik itu dari dokumen yang diunggah atau teks yang dibuat untuk penggunaan online. Batasan yang lebih lembut adalah bahwa mengandalkan sepenuhnya pada tautan yang ditemukan pada subjek Reddit, kumpulan data untuk mewarisi karakteristik dan bias Reddit sebagai platform, seperti demografi penggunanya dan tingkat perputaran konten yang tinggi. Survei Pew Internet Research 2013 mengungkapkan pengguna Reddit di Amerika Serikat lebih cenderung laki-laki, di akhir usia belasan hingga dua puluhan, dan urban [16]. Jadi, kumpulan data terdiri dari bahan-bahan yang memiliki relevansi topik dengan diskusi online di antara para demografis ini. Kami mengklasifikasikan pendekatan scraping web ini sebagai *laissez-faire,* dengan satu-satunya bentuk intervensi yang menyaring tautan jika pos memiliki kurang dari tiga suara positif bersih, dan menggunakan daftar blokir untuk menghindari subreddits yang berisi "konten seksual eksplisit atau menyinggung" [11]. Mengingat banyak penelitian yang menunjukkan keraguan atas suara positif sebagai indikasi popularitas atau kualitas karena faktor-faktor seperti bias "perantara", suara bot, dan suara palsu, mekanisme penyaringan ini dapat dianggap sewenang-wenang [22, 23, 33]. Penulis juga menghapus semua konten Wikipedia untuk mengurangi kerumitan dengan validasi model.

Kami dapat meningkatkan transparansi dan secara eksplisit menggambarkan batasan inklusivitas sosiokultural dengan menyertai WebText dengan
Pernyataan Misi yang menjelaskan ruang lingkup materi serta tujuan pengumpulan. Contohnya adalah:

> Pernyataan Misi WebText WebText adalah kumpulan data salinan web dari dokumen online yang ditautkan dari Reddit.com. Saat ini terdiri dari ~ 40G teks dari 8 juta dokumen. Motivasinya adalah untuk mengumpulkan 1. sejumlah besar data bahasa alami 2. di berbagai konteks dan domain untuk mengoptimalkan kinerja GPT-2, model bahasa yang dilatih untuk ditransfer ke berbagai tugas NLP. WebText disusun untuk tujuan penelitian ini bereksperimen dengan model bahasa, bukan untuk penggunaan komersial. Selain menyaring tautan dengan kurang dari 3 suara positif atau tautan ke artikel Wikipedia, WebText sepenuhnya non-intervensionis. WebText bertujuan untuk meningkatkan ukuran kumpulan data dengan terus mengorek halaman yang diperbarui di Reddit.

Pernyataan Misi tersebut menguraikan apa asal mula Teks Web, tingkat intervensi yang diterapkan dalam pembuatannya, dan aplikasi apa yang lebih sesuai untuk itu. Selanjutnya, jika WebText bersifat publik, bagian dari a konsorsium data, atau menerima kontribusi eksternal, ini bisa menjadi sinyal bagaimana pihak lain dapat berkontribusi pada kumpulan data khusus ini.

Namun, jika WebText, atau kumpulan data besar lainnya dengan tujuan melatih model bahasa, dimaksudkan untuk melatih model bahasa Inggris yang komprehensif, pendekatan yang lebih intervensionis mungkin sesuai bersama dengan pernyataan misi yang dimodifikasi. Contoh Pernyataan Misi yang menargetkan bahasa Inggris Amerika non-internet untuk melengkapi pendekatan yang berpusat pada Reddit adalah di bawah ini:

> Lengkapi DatasetMissionStatement Misi kumpulan data ini adalah mengumpulkan bahasa dari kelompok-kelompok di Amerika Serikat yang tidak secara teratur berkontribusi pada bahasa Inggris internet. Banyak model NLP melatih data bahasa Inggris yang ditemukan di internet, melalui situs penyiaran teks alami yang besar seperti Wikipedia, Reddit, serta situs ulasan film dan restoran seperti Yelp dan IMDB. Namun, bahasa Inggris 'bahasa alami' di Amerika Serikat dituturkan oleh demografi dalam jangkauan yang lebih luas daripada yang disajikan di internet. Kumpulan data ini bertujuan untuk secara aktif mengumpulkan keragaman bahasa Inggris yang dituturkan oleh budaya Amerika secara luas untuk berkontribusi pada sistem ML seperti model bahasa sehingga mereka dapat menggabungkan ekspresi, topik minat, keyakinan, dan struktur tata bahasa dari orang-orang yang kurang terwakili di internet. Dataset ini terutama difokuskan pada bahasa Inggris sehari-hari di seluruh kelas,

SEBUAH kebijakan pengembangan koleksi dapat dikembangkan secara lebih sistematis untuk mengatasi kesenjangan dalam keragaman dan inklusivitas sosiokultural. Di sinilah rencana tingkat tinggi hingga menengah dibuat tentang bagaimana mengejar Pernyataan Misi dan bagaimana mengumpulkannya. Di sini kami dapat menyusun daftar demografi target yang data linguistiknya sedikit digital atau mungkin ditemukan di luar relevansi Reddit:

Beberapa deskripsi dasar Redditers dan batasan keras WebText yang semua dokumennya ditemukan online memberi kami demografi target dasar [16]:

- Varians Generasi: Menurut statistik PEW, kebanyakan redder mewakili generasi milenial hingga pasca milenial. Suatu pendekatan untuk melengkapi WebText adalah dengan mengidentifikasi di mana data bahasa atau topik yang relevan dengan generasi lain dapat ditemukan.
- Jenis kelamin: Redditers dua kali lebih mungkin menjadi laki-laki daripada perempuan. Kami dapat menyeimbangkan konten dengan memeriksa gender dari tautan di Reddit.

- Akses Internet: Penduduk daerah dengan akses internet terbatas cenderung tidak berkontribusi pada diskusi Reddit. Salah satu strategi untuk melengkapi adalah dengan mengidentifikasi wilayah Amerika Serikat di mana akses internet terbatas atau penggunaan rendah.
- Pedesaan Amerika: Redditers lebih cenderung menjadi urban. Ini dapat merusak distribusi topik relevansi. Pedesaan Amerika lebih cenderung memiliki minat dan kecenderungan politik yang berbeda. Misalnya, pedesaan Amerika lebih cenderung memilih Republikan daripada di situs perkotaan.

- Imigrasi: Lingkungan multi-etnis / imigran di mana bahasa Inggris bukan bahasa dominan atau penduduk mengonsumsi budaya asing karena sumber yang dominan dapat menjadi wilayah lain di mana bahasa Inggris non-tandard dapat dikumpulkan.

Kami kemudian dapat mengejar strategi pengumpulan data seperti: (1) Arsip Komunitas: Identifikasi arsip komunitas yang menyimpan materi yang relevan dengan demografi yang diminati. Apakah ada arsip komunitas di seluruh negeri yang telah mengumpulkan bahan-bahan bahasa?

(2) Skema Partisipatif: Hasilkan arsip partisipatif. Kirim

keluarkan peralatan lengkap atau staf ke komunitas pedesaan, lingkungan multi-etnis, dan area dengan penggunaan internet rendah. Kumpulkan sejarah lisan dan wawancara berdasarkan koleksi bahasa yang mencakup berbagai topik.

Untuk memastikan pengawasan etis dalam pengumpulan data, proyek perlu mempekerjakan waktu penuh staf pengumpulan dan pengelolaan data dengan keanggotaan dalam organisasi ekstra-arsip dan menyusun atau mengadopsi a Kode etik untuk proyeknya. Ini memastikan staf yang terlibat dalam pengumpulan data dimintai pertanggungjawaban atas pelanggaran etika. WebText saat ini hampir sepenuhnya *non-intervensionis,* tanpa penyaringan untuk data pribadi. Tomake mempertimbangkan keputusan yang menghitung keuntungan dan risiko penambahan dan pengurangan subset data, yang dapat diselenggarakan oleh proyek komite termasuk ahli domain. Komite dapat melakukan intervensi dengan memasukkan data sensitif, data ekstremis (misalnya ekstremisme politik), dan data penargetan ras / gender.

Akhirnya, pelepasan mengiringi dokumentasi pada komposisi WebText dapat meningkatkan transparansi konten yang menggerakkan model. Saat ini tidak ada penyebutan eksplisit tentang proses pemilihan - jika ada - subreddits yang digunakan, jam berapa data dikikis, mengaburkan konten. Jika lebih banyak metode intervensionis diterapkan, tahapan pengambilan keputusan ini dapat dicatat untuk lebih meningkatkan transparansi.

8 BATASAN

Ada beberapa peringatan penting untuk proposal yang terinspirasi oleh metode pengumpulan data arsip. Set data ML cenderung lebih besar dan studi tambahan tentang penskalaan pedoman ini akan menginformasikan bagaimana data tersebut dapat ditransfer dalam konteks ML. Secara khusus, mempekerjakan staf penuh waktu, menyimpan dokumentasi, dan menerapkan strategi pengumpulan dalam skala besar menimbulkan biaya overhead, waktu, dan keuangan yang besar. Mempertahankan konsorsium data di tingkat komunitas adalah salah satu cara untuk mengurangi biaya ini dengan skala ekonomi, berbagi sumber daya, dan meminimalkan duplikasi. Upaya ini membutuhkan koordinasi yang substansial dan upaya masyarakat. Tetapi seperti yang telah kita lihat, institusi besar mengadopsi kerangka kerja audit dan dokumentasi seperti Lembar Data, Kartu Model, dan Lembar Fakta, investasi ini bukan tidak mungkin.

Kumpulan data dan arsip ML juga memiliki perbedaan intrinsik dalam motivasi. Banyak proyek ML dimotivasi secara komersial dan didanai perusahaan, mengarahkan fokus pada pengguna internet dan konsumen teknologi dan insentif untuk menjaga kepemilikan data. Arsip dan perpustakaan bertujuan untuk melestarikan warisan budaya dan keragaman. Analisis model bisnis adalah diskusi penelitian sendiri yang diperlukan, di luar cakupan makalah ini. Pertimbangan lebih lanjut tentang model insentif dalam ML dapat menjelaskan bagaimana komunitas dapat menyesuaikan rekomendasi dari arsip ini.

Akhirnya, model intervensionis bukannya tanpa kesalahan. Satu perhatian adalah bahwa pendekatan pengumpulan data yang sangat selektif memusatkan kekuasaan pada arsiparis dalam menentukan portofolio dan memperlakukan materi secara etis. Pengaruh sosial dan politik yang tidak semestinya untuk mengatur agenda adalah hal lain. Namun, sistem intervensi multi-lapis dan multi-orang masih menyebarkan kekuatan di antara lebih banyak orang dan lebih sistematis daripada ketika seorang insinyur ML mengompilasi set data. Badan pemerintah di dalam dan di seluruh lembaga untuk mengaudit kolektor juga menyediakan langkah-langkah untuk meminta pertanggungjawaban kolektor. Tidak satu pun dari pengamanan ini yang saat ini diterapkan dalam komunitas ML.

9 KESIMPULAN

Makalah ini telah menunjukkan bagaimana arsip dan perpustakaan, bidang yang didedikasikan untuk pengumpulan data manusia untuk anak cucu telah bergulat dengan pertanyaan etika, representasi, kekuasaan, transparansi, dan persetujuan. Strategi ini bersifat institusional dan prosedural, membutuhkan dana yang dialokasikan dan upaya kolektif lembaga besar dan kecil. ML menghadirkan tantangan tambahan dalam menangani audiens yang lebih luas dan mendorong produk komersial. Dan meskipun banyak arsip bersifat nirlaba dan pendidikan, kumpulan data ML sering kali dikaitkan dengan tujuan laba atau pertahanan, meningkatkan taruhan pengumpulan data yang bermasalah. Dengan demikian, arsip investasi yang telah dibuat dalam praktik pengumpulan data etis sangat mungkin dilakukan dalam ML.

Arsip bukanlah satu-satunya tempat yang dapat kita pelajari. Untuk berurusan dengan subjek manusia langsung, dan masalah privasi dan representasi, kita dapat menarik dari ilmu sosial eksperimental dan kerja lapangan seperti sosiologi dan psikologi. Sejarawan berpengalaman dalam konteks sejarah dan antropolog dalam kepekaan budaya. Dalam menavigasi jalur yang belum dipetakan, komunitas ML dapat melihat bidang yang lebih lama untuk contoh keberhasilan dan kegagalan pada hal-hal yang sebanding.

REFERENSI

[1] Imran Ahmed, Giles L Colclough, Daniel First, dkk. 2019. Mengoperasikan Manajemen Risiko untuk Pembelajaran Mesin: Membangun Sistem Berbasis Protokol untuk Kinerja, Penjelasan, & Keadilan. (2019).

[2] ALA. 2008. *Kode Etik American Library Association.* Perpustakaan Amerika Asosiasi. http://www.ala.org/tools/ethics

[3] ARA. 2018. *Kode etik.* Asosiasi Arsip & Arsip (Inggris & Irlandia). https://www.archives.org.uk/membership/code-of-ethics.html

[4] Bank Dunia. 2018. Orang Perorangan yang Menggunakan Internet. (2018). https: //data.worldbank. org / indikator / IT.NET.USER.ZS? end = 2017 & lokasi = AS & mulai = 2015

[5] Blog Google Australia. 2018. Mempromosikan kebanggaan dan rasa hormat dengan kecerdasan buatan gence. https://australia.googleblog.com/2018/02/promoting-pride-and-respect- with.html

[6] Richard Breitman dan Norman JW Goda. 2010. *Bayangan Hitler: Perang Nazi Penjahat, Intelijen AS, dan Perang Dingin.* Arsip Nasional. [7] Caroline Brown. 2014. *Arsip dan Pencatatan: Teori menjadi Praktik.* Segi Penerbitan.

[8] Joy Buolamwini dan Timnit Gebru. 2018. Nuansa gender: Akurasi titik-temu ketimpangan dalam klasifikasi gender komersial. Di *Konferensi tentang Keadilan, Akuntabilitas dan Transparansi.* 77–91.

[9] Pusat Pelestarian Budaya Catawba. 2019. Angsa Tradisional Wanita Menari. http://catawbaarchives.libraries.wsu.edu/digital-heritage/womens-tradisional-angsa-tari-3 Koleksi Wenonah G. Haire.

[10] Peter Cihon. 2019. *Standar untuk Tata Kelola AI: Standar Internasional untuk Memungkinkan Koordinasi Global dalam Riset & Pengembangan AI.* Universitas Oxford. https: // www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf [11] Jack Clark. 2019. Kartu Model GPT-2. gpt2-modelcardshttps: //github.com/openai/ gpt-2 / blob / master / model_card.md.

[12] Maygene F. Daniels dan Timothy Walch. 1984. *Pembaca Arsip Modern: Dasar Bacaan tentang Teori dan Praktek Kearsipan.* Layanan Arsip dan Arsip Nasional.

[13] Elena S. Danielson. 2004. Hak Privasi dan Hak Korban Politik Implikasi dari Pengalaman Jerman. *Pengarsip Amerika* 67 (2004), 176–193. https://americanarchivist.org/doi/pdf/10.17723/aarc.67.2.1w06730777226771

[14] Denise M. Davis. 2006. *Jaringan Perpustakaan, Koperasi, dan Konsorsium: Defi-Studi dan Survei Nasional.* http://www.ala.org/aboutala/sites/ala.org.aboutala/ files / content / ors / lncc / interim_report_1_may2006.pdf

[15] Departemen Media & Olahraga untuk Digital, Budaya. 2018. *Pusat Etika Data dan Inovasi: Konsultasi.* https://assets.publishing.service.gov.uk/government/ uploads / system / uploads / attachment_data / file / 715760 / CDEI_consultation__1_ . pdf

[16] Maeve Duggan dan Aaron Smith. 2013. 6% orang dewasa daring adalah pengguna reddit. *Bangku gereja Proyek Internet & Kehidupan Amerika* 3 (2013), 1–10.

[17] Andrew Flinn. 2007. Sejarah Komunitas, Arsip Komunitas: Beberapa Peluang dan Tantangan. *Jurnal Society of Archivists* 28, 2 (2007), 151–176. https://doi.org/10.1080/00379810701611936 arXiv: https: //doi.org/10.1080/00379810701611936

[18] Nancy Freeman dan Holly Geist. 2014. Permintaan FOIA. *Studi Kasus di Arsip Etika (* 2014).

[19] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, dan James Zou. 2018. Kata embeddings mengukur 100 tahun gender dan stereotip etnis. *Prosiding Akademi Ilmu Pengetahuan Nasional* 115, 16 (2018), E3635 – E3644.

[20] Clare Garvie, Bedoya Alvaro, dan Jonathan Frankle. 2016. *Line-up abadi: Pengenalan wajah polisi yang tidak diatur di Amerika.* Hukum Georgetown, Pusatkan Privasi & Teknologi.

[21] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, HannaWallach, Hal Daumeé III, dan Kate Crawford. 2018. Lembar data untuk kumpulan data. *arXiv pracetak arXiv: 1803.09010 (* 2018).

[22] Eric Gilbert. 2013. Penyebaran yang Luas di Reddit. Di *Prosiding dari Konferensi 2013 tentang Kerja Koperasi yang Didukung Komputer (CSCW '13).* ACM, Baru York, NY, AS, 803–808. https://doi.org/10.1145/2441776.2441866

[23] M. Glenski, C. Pennycuff, dan T. Weninger. 2017. Konsumen dan Kurator: Browsing dan Pola Voting di Reddit. *Transaksi IEEE pada Sistem Sosial Komputasi* 4, 4 (12 2017), 196–206. https://doi.org/10.1109/TCSS.2017.2742242

[24] Tracy B. Grimm dan Chon A. Noriega. 2013. Mendokumentasikan Seni Latin Daerah dan Budaya: Studi Kasus untuk Pendekatan Kolaboratif Berorientasi Komunitas. *Pengarsip Amerika* 76 (2013), 95–112. [25] F. Gerald Ham. 1993. *Memilih dan Menilai Arsip dan Naskah.* Itu

Masyarakat Arsiparis Amerika.

[26] Foad Hamidi, Morgan Klaus Scheuerman, dan Stacy M. Branham. 2018. Jenis Kelamin Pengakuan atau Reduksionisme Gender ?: Implikasi Sosial dari Sistem Pengakuan Gender yang Tertanam. 1–13. https://doi.org/10.1145/3173574.3173582 [27] Ailsa C. Holland dan Elizabeth Mullins. 2013. *Arsip dan Pengarsip 2: Saat ini Tren, Suara Baru.* Empat Pengadilan Pers.

[28] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, III, Miro Dudik, dan Hanna Wallach. 2019. Meningkatkan Keadilan dalam Sistem Pembelajaran Mesin: Apa Apakah Praktisi Industri Perlu ?. Di *Prosiding Konferensi CHI 2019 tentang Faktor Manusia dalam Sistem Komputasi (CHI '19).* ACM, New York, NY, AS, Artikel 600, 16 halaman. https://doi.org/10.1145/3290605.3300830

[29] Valerie Horton dan Greg Pronevitz. 2015. *Perwakilan Perpustakaan: Model untuk Kolaborasi ransum dan Keberlanjutan.* Edisi ALA.

[30] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, dan Ralph Weischedel. 2006. OntoNotes: Solusi 90%. Di *Prosiding Manusia Konferensi Teknologi Bahasa NAACL, Volume Pendamping: Makalah Singkat (NAACL-Short '06).* Asosiasi Linguistik Komputasi, Stroudsburg, PA, AS, 57-60. http://dl.acm.org/citation.cfm?id=1614049.1614064 [31] ICRM. 2016. *Kode etik.* Institut Manajer Arsip Bersertifikat. https: //www.icrm.org/code-of-ethics

[32] Lembaga Data Terbuka. 2019. Laporan Ringkasan Data Trusts. https://theodi.org/wp-content / uploads / 2019/04 / ODI-Data-Trusts-A4-Report-web-version.pdf [33] Pascal J *u* Ü rgens dan Birgit Stark. 2017. Kekuatan Default pada Merah-dit: Model Umum untuk Mengukur Pengaruh Perantara Informasi. *Kebijakan & Internet* 9, 4 (2017), 395–419. https://doi.org/10.1002/poi3.166 arXiv: https: //onlinelibrary.wiley.com/doi/pdf/10.1002/poi3.166

[34] Randall C. Jimerson. 2009. *Kekuatan arsip: ingatan, akuntabilitas, dan sosial keadilan.* Chicago: Perkumpulan Arsiparis Amerika.

[35] Anna Jobin, Marcello Ienca, dan Effy Vayena. 2019. Lanskap global AI pedoman etika. *Kecerdasan Mesin Alam* 1, 9 (2019), 389–399.

[36] James E Johndrow, Kristian Lum, dkk. 2019. Algoritme untuk menghapus sensitif informasi: aplikasi untuk prediksi residivisme ras-independen. *The Annals Statistik Terapan* 13, 1 (2019), 189–220.

[37] Os Keyes. 2019. Counting the Countless: Mengapa data science adalah ancaman yang sangat besar untuk orang aneh. (2019). https://reallifemag.com/counting-the-countless/

[38] Os Keyes, Nikki Stevens, dan Jacqueline Wernimont. 2019. Pemerintah Apakah Menggunakan Orang Paling Rentan untuk Menguji Perangkat Lunak Pengenalan Wajah. *Batu tulis Majalah (* 2019).

[39] Christopher Kitching dan Ian Hart. 1995. Pernyataan Kebijakan Koleksi. *Jour-nal dari Society of Archivists* 16 (1995). http://www.nationalarchives.gov.uk/ documents / archives / archive-collection-policy.pdf

[40] Philipp Koehn. 2005. Europarl: Sebuah korpus paralel untuk terjemahan mesin statistik tion. Di *KTT MT,* Vol. 5. 79–86. [41] Ivana KottasovÃą. 2018. Perusahaan-perusahaan ini terbunuh oleh GDPR. (2018). https://money.cnn.com/2018/05/11/technology/gdpr-tech-companies- losers / index.html

[42] David Lowenthal. 2006. Archival Perils: An Historian's Pliant. *Arsip: The Jurnal Asosiasi Rekaman Inggris (* 2006), 49–75. https://doi.org/10.3828/ arsip. 2006.6

[43] Heather MacNeil. 1992. *Tanpa Persetujuan: Etika Mengungkapkan Informasi Pribadi mation di Arsip Umum.* The Scarecrow Press, Inc.

[44] Majalah Life. 1941. Bagaimana Mengenalinya Orang Jepang dari Orang Cina. *Majalah Life* 11, 25 (12 1941), 81. http://digitalexhibits.wsulibs.wsu.edu/items/show/4416 AP2.L547.1941.12.22.p81.

[45] Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, dan Britta Schasberger. 1994. The Penn Tree- bank: Annotating Predicate Argument Structure. Di *Prosiding Lokakarya Teknologi Bahasa Manusia (HLT '94).* Association for Computational Linguis- tics, Stroudsburg, PA, USA, 114–119. https://doi.org/10.3115/1075812.1075835

[46] Katherine McCardwell. 2014. Kekhawatiran Kekayaan Intelektual di Tidak Berdokumen Koleksi Perusahaan. *Studi Kasus dalam Etika Kearsipan (* 2014).

[47] Margaret Mitchell, SimoneWu, AndrewZaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, dan Timnit Gebru. 2019. Kartu model untuk pelaporan model. Di *Prosiding Konferensi tentang Keadilan, Akuntabilitas, dan Transparansi.* ACM, 220–229.

[48] Paul Mozur. 2019. Satu Bulan, 500.000 Pemindaian Wajah: Bagaimana China Menggunakan AI Buat Profil Minoritas. *The New York Times (* 2019).

[49] MS. 1999. *The Manuscript Society of Code of Ethics Diadopsi oleh Dewan Pengawas 26 Mei 1999.* The Manuscript Society. manuscript.org/about/code-of-ethics [50] Institut Museum dan Layanan Perpustakaan. 2018. imls.gov/sites/default/files/ grants / sp-02-16-0015-16 / proposals / sp-02-16-0015-16_proposal_documents. pdf

[51] SEPASANG Google. 2020. Buku Panduan Orang + AI. https://pair.withgoogle.com

[52] Pew. 2018. Lembar Fakta Internet / Broadband. (2 2018). https: //www.pewinternet. org / lembar fakta / internet-broadband /

[53] Timothy D. Pyatt. 2015. The Harding Affair Letters: How One Archivist Took Setiap Tindakan Mungkin Untuk Memastikan Pelestariannya. *Studi Kasus dalam Etika Kearsipan (* 2015).

[54] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, dan Ilya Sutskever. 2019. Model bahasa adalah pelajar multitask tanpa pengawasan. *Blog OpenAI* 1, 8 (2019).

[55] Ellen M. Ryan. 2014. Mengidentifikasi Bahan Indian Amerika yang Sensitif Secara Budaya di Lembaga Non-suku. *Studi Kasus dalam Etika Kearsipan (* 2014).

[56] SAA. 2011. *Pernyataan Nilai Inti SAA dan Kode Etik.* Masyarakat Amerika Arsiparis. https://www2.archivists.org/statements/saa-core-values-statement- dan-kode-etik

[57] SAA. 2013. *Menjelaskan Arsip: Standar Isi (* edisi kedua). Masyarakat Arsiparis Amerika.

[58] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, dan Janet Vertesi. 2019. Keadilan dan abstraksi dalam sistem sosioteknik. Di *Prosiding Konferensi tentang Keadilan, Akuntabilitas, dan Transparansi.* ACM, 59–68.

[59] Olivia Solon. 2019. 'Rahasia Kecil Kotor' Pengenalan Wajah: Mil-Singa Foto Online yang Diambil tanpa Persetujuan. *NBC News (* 2019). https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-rahasia-jutaan-online-foto-scraped-n981921

[60] Harini Suresh dan John V. Guttag. 2019. Kerangka untuk Memahami Unin-Konsekuensi Pembelajaran Mesin yang cenderung. *ArXiv* abs / 1901.10002 (2019). [61] Aditya Tripathi dan Jawahar Lal. 2016. *Konsorsium Perpustakaan: Panduan Praktis untuk Manajer Perpustakaan.* Penerbitan Chandos.