Current Topological and Machine Learning Applications for Bias Detection in Text

Colleen Farrelly*, Yashbir Singh†, Quincy A. Hathaway‡, Gunnar Carlsson§, Ashok Choudhary¶, Rahul Paul‖
Gianfranco Doretto‡ Yassine Himeur** Shadi Atalls** Wathiq mansoor**
* Staticlysm LLC, Miami, FL, USA
† Radiology, Mayo Clinic, Rochester, MN
‡ West Virginia University, Morgantown, WV, USA
§Stanford University, California, USA
¶Department of surgery, Mayo Clinic, Rochester, MN, USA
‖Harvard University, USA
**College of Engineering and Information Technology, University of Dubai, Dubai, UAE

2023 6th International Conference on Signal Processing and Information Security (ICSPIS) | 979-8-3503-2959-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSPIS60075.2023.10343824

***Abstract*—Institutional bias can impact patient outcomes, educational attainment, and legal system navigation. Written records often reflect bias, and once bias is identified; it is possible to refer individuals for training to reduce bias. Many machine learning tools exist to explore text data and create predictive models that can search written records to identify real-time bias. However, few previous studies investigate large language model embeddings and geometric models of biased text data to understand geometry's impact on bias modeling accuracy. To overcome this issue, this study utilizes the RedditBias database to analyze textual biases. Four transformer models, including BERT and RoBERTa variants, were explored. Post-embedding, t-SNE allowed two-dimensional visualization of data. KNN classifiers differentiated bias types, with lower k-values proving more effective. Findings suggest BERT, particularly mini BERT, excels in bias classification, while multilingual models lag. The recommendation emphasizes refining monolingual models and exploring domain-specific biases.**

***Index Terms*—Topological data analysis, machine learning, Natural language processing, Bias, Text Embeddings.**

I. Introduction

In recent years, the intersection of topological data analysis and machine learning has opened up exciting new avenues for understanding and addressing the issue of bias in text data [1]–[3]. With the proliferation of textual information on the internet, the potential for bias, both subtle and overt, has become a pressing concern in natural language processing (NLP) and text analysis [4], [5]. Bias can manifest in various forms, such as gender bias, racial bias, or political bias, and can significantly impact the fairness and accuracy of text-based systems, including search engines, recommendation systems, and sentiment analysis tools [6]–[8].

A. Stigma and Bias

Stigma is a perceived identity within a society or subgroup arising from a mismatch between social identities valued by those in power within a society and a person's true identity [9]. Examples include physical characteristics (such as race or missing limb differences), medical disorders (such as substance use disorders or visual impairment), and changeable visual cues of subgroups (such as tattoos or mohawks within a society where the majority and those in power do not display these visual cues). Stigma can be associated with two forms of bias: 1) implicit biases, where the person within a majority group acts with bias but is unaware that they are acting on their bias against the stigmatized group [10], and 2) explicit biases, where the person within the majority group acts with bias and is aware that they are acting on their bias against the stigmatized group [11]. Thus, stigmatized groups often face unique societal challengesbecause of implicit and explicit biases, whichcan happen formally or informally.

B. Institutional Bias

Characteristics devalued by those in power or by the majority in society as different than the society's norms become codified into laws and regulations within societal institutions, and individuals with those characteristics often face institutional barriers [9]. For instance, a student with visual impairment in an educational system that assumes all students can see will run into problems with many academic tasks, like reading assignments or the blackboard, unless another option is created; fortunately, this is usually the result of an implicit bias that is quickly identified by the educational system [12]. Jim Crow laws in the American South exemplify explicit bias within the legal system [13].

C. Non-Institutional Bias

Devaluation of individuals based on characteristics can also be more informal with respect to everyday interactions. For instance, a physician treating a patient with a history of substance use disorder which presents with pain may quickly label the patient as "drug-seeking," which then follows the patient throughout their journey to receive a diagnosis [14], [15]; hopefully, this is an implicit bias and not an intentional dismissal of a patient. Implicit bias against female patients has been well-documented as leading to adverse outcomes for patients presenting in emergency settings with symptoms that diverge from androcentric cardiac symptoms in emergency settings [16]–[18]. Both types of informal biases often become codified in the clinical record by physicians, creating institutional barriers through documented language by those in authority [19]–[22]

D. Implicit Bias Solutions

Because implicit bias is not intentional, antibias training work effectively in combatting bias and have lasting effects [23], [24]. Within the medical setting, implicit biases against patients can be ameliorated through diversity training, such as in Morris et al., 2019, which provided a curriculum for working with LGBTQIA+ patients for current students. Interventions like this can likely assuage the issues in other fields, such as the legal field or education. Creating inclusive environments has also been shown to combat biases [25].

E. Natural Language Processing

One branch of machine learning, natural language processing (NLP), processes and analyzes text data [26]. Many tools exist in this field, but within the context of bias detection, a few tools are more relevant. Typically, a block of text is first parsed into individual words via a process called tokenization [27], [28]. Then, each token can be matched to terms of interest, through a pre-trained algorithm that recognizes terms of interest (such as famous people or places) or a custom matching dictionary. However, for large document sets with an extensive list of terms for which to search, this can be computationally intensive [4], [29]. Another tool commonly used to wrangle text data into matrix form for use in supervised learning models, such as logistic regression or random forest classifiers, is embedding text. Many ways to do this exist. First, one can simply encode the full sets of documents by frequency of each term that exists at least once in the set of documents, usually according to preset transformations or weightings of frequencies across words and across documents [30]; term frequency–inverse document frequency (TF-IDF) is one common method to transform text documents [30]. Many other embeddings exist, including word2vec and GloVe [31], [32]. One drawback of this method is the potential dimensionality of the matrix, which can become problematic for statistical models or machine learning algorithms modeling the data from the embedding matrix. Another embedding option is to use a pre-trained model that can embed text according to the text meaning and map these meanings to a matrix with a pre-defined size to limit the potential size of the matrix after embedding all the text documents. Bidirectional Encoder Representations from Transformers (BERT) is one such option; implementations of BERT typically use an algorithm to map new text to the pre-trained embedding [33]. This process can be computationally intensive, though, and it can require additional computing resources to obtain the full document set embedding; further, matrices typically don't have as many potential predictors as a TF-IDF embedding. BERT has already been used to detect online hate speech [34], [35]. In addition, some studies have shown racial and gender bias in word embeddings themselves, so such a method may not be suited for bias detection applications [36]. In addition, hundreds of BERT variations exist, and it is not known how different variations perform with respect to bias detection.

F. Machine Learning on Text Embeddings

Once text data is embedded through TF-IDF, BERT, or other methods, machine learning, and statistical models can be used to explore and predict biases within the text documents. Supervised learning, where the machine learning model learns to predict a particular outcome (such as spam/not spam or biased/not biased), is a common task in text analytics [37], [38]. Models are trained on a set of embedded text documents, followed by validation on the remaining text documents to ensure the model can generalize to other documents. Many studies have examined ways to classify racial or gender bias within document sets [34], [35], [39]. k-nearest neighbor (KNN) classifiers are common in text classification problems and rely on local geometry to classify a point based on the points nearest that point in the geometry of the embedding space [40]–[43]. For the classification of technical documents, Larkey and Croft (1996) found advantages of KNN models and embedding strategies; given the linguistic nuances of biased language, it is possible that embedding strategies coupled with KNN models will work well for bias detection. Unsupervised learning, which includes data mining methods like clustering or visualization, can also be useful in text bias detection, particularly when specific terms are not known ahead of time. Unsupervised learning on text data has been used to understand what exists in qualitative data [44], to explore topics in text datasets, and to summarize texts [45], among many other applications. Unsupervised learning may be a good first step in bias detection pipelines, as it can uncover bias types in clinical notes. T-stochastic neighbor embedding (t-SNE), a dimensionality reduction technique that creates pairwise probability distributions to embed points in lower-dimensional spaces, has been used in text visualization [46], [47]. Given t-SNE's ability to visualize high-dimensional text embeddings in low-dimensional space, it is likely that t-SNE will make a good tool for exploring bias separation in text embeddings such as BERT.

II. Methods

Data consisted of text samples curated in the RedditBias database (https://github.com/SoumyaBarikeri/RedditBias), containing religious, racial, gender, and orientation bias [48]. Because religious bias had a larger text sample (2139 for religious. 504 for each of the others) than other bias types, we employed random sampling on the religious bias dataset to create a similar-sized subsample of religious bias (504 examples) compared to other bias subsamples (504 examples each in racial, gender, and orientation bias).

Four types of pre-trained transformer models were fit on the RedditBias samples. Our first BERT model was the all-mpnet-base-v2 (full BERT) model from the HuggingFace repository of Python's sentence_transformer package [49], which is a sentence embedding version of BERT trained on 1 billion sentence pairs. Our second BERT model was all-MiniLM-L6-v2 model (mini BERT), which uses the same training set as our full BERT model but embeds the data into a smaller space (creating a dense embedding). Our first RoBERTa model was the all-roberta-large-v1 model (all RoBERTa), which uses the same training data with the RoBERTta design rather than BERT design of transformer and a sparser, high-dimensional embedding space. Our second RoBERTa model was the xlm-roberta-base model (raw RoBERTa), which trains on 2.5 terabytes of scraped data across 100 languages and embeds result in a smaller space with a denser embedding without any data curation or validation prior to training the model.

After embedding the data, t-SNE was applied to the embeddings to reduce dimensionality to two dimensions for easy visualization. As mentioned, t-SNE is a dimensionality reduction method that relies on pairwise probability distribution calculation to embed pairs of points into a lower-dimensional space [50] and falls into the broad field of dimensionality reduction techniques. We implemented t-SNE (Default parameters of TSNE from sklearn manifold package with n_components=2) on our four embeddings through scikit-learn's Python implementation of t-SNE [51]. KNN classifiers classify point labels based on the k points nearest a given point through label voting [52]. If k=5 and four points are label X while one point is label Y, the point in question would be assigned the label X, as more votes from neighbors were for the label X. Thus, local geometry and point distributions play a large role in the performance of KNN models. Distance metric choice can play a large role in performance [53]. The choice of k can lead to either underfitting (larger k values) or overfitting (smaller k values) of a model, with k acting as a smoothing parameter [54]. We fit our KNN classifiers with sci-kit-learn's Python implementation of t-SNE, a varying choice of k (3, 10, or 25), a Euclidean distance metric (which is native to BERT and RoBERTa embeddings), and the four embeddings themselves as algorithm input (rather than their reduced forms in the t-SNE application) (Fig. 1). KNN classifiers were run 50 times each to compare results. Bonferroni-corrected t-tests then served to distinguish classifier performance statistically across runs.

III. Results

The t-SNE plots suggest that most of our chosen embeddings separate bias types well (Fig. 2, Fig. 3, Fig. 4, and Fig. 5). The raw text RoBERTa model that included many language examples did not perform as well as the other three embeddings, suggesting that a manual review of training data and focus on embeddings that only include the language of interest may be useful in the context of bias detection. BERT embeddings may perform better than RoBERTa embeddings in this context, as well. For the best embeddings, such as our full BERT embedding, almost no overlap exists, suggesting that KNN models would perform very well on the classification task.

Models with k=3 and k=10 performed better across embeddings (70% training, 30% test)than k=25 ($p<0.01$), and BERT embeddings showed superior performance to RoBERTa embeddings across choices of k ($p<0.01$), with the mini BERT embedding showing the best overall performance, particularly at k=3 ($p<0.01$) (Table I).

Table I. Embedding type and number of nearest neighbors impacted KNN classifier performance. Our sample data included misspellings, grammatical mistakes, and varying text lengths. These results suggest that our chosen BERT models and the single-language RoBERTa model are robust to real-world English-language text, including text with misspellings and grammatical errors. However, multilingual models without validation or cleaning of training data do not seem to work as well, though they are currently the only option for multilingual document sets.

TABLE I: Embedding Results

Embedding	$K = 3$	$K = 10$	$K = 25$
Full BERT	$0.99(0.98, 1.00)$	$0.99(0.97, 1.00)$	$0.99(0.97, 1.00)$
Mini BERT	$1.00(0.99, 1.00)$	$0.99(0.98, 1.00)$	$0.99(0.98, 1.00)$
Full RoBERTa	$0.99(0.98, 1.00)$	$0.99(0.98, 1.00)$	$0.99(0.97, 0.99)$
Raw RoBERTa	$0.87(0.84, 0.90)$	$0.88(0.82, 0.91)$	$0.88(0.83, 0.91)$

IV. Discussion

Logical next steps include pilot studies of applying this methodology to educational notes, medical notes, and legal notes to assess the accuracy and feasibility of detecting bias in these sources based on the proposed embedding strategy coupled with KNN models. It is possible that embeddings will not work as well in fields with substantial jargon, such as the medical field, and testing field-specific embeddings may be worth trying. Further, the medical field, including the electronic medical record and clinical notes, likely has lower frequencies of overt racial, religious, gender, and orientation bias. Applying our proposed model within healthcare likely would include tailoring the algorithm to identify more difficult forms of bias, such as negative patient descriptors [11]; these could include terms such as aggressive, agitated, angry, challenging, combative, etc.

In addition, the assessment of value-added and the potential human impact of applying these classifiers to real-world institutional data should be a part of these pilot studies. However, given the promising results of this study, it is worth exploring this methodology on field-specific text data. Reliable detection of different types of bias can pinpoint areas of improvement for institutions and individuals and pinpoint what type of training would be needed to reduce bias. Reduced institutional bias can, in turn, improve outcomes for those interacting with institutions.

Because our data involves only Reddit (Reddit Inc © 2023) samples, it is hard to know how the results would generalize to longer text documents or to text documents with field-specific jargon. Because prior studies have shown that good embeddings plus a KNN model can improve text classifier accuracy [29], it is likely that the methodology will generalize, if not the specific embeddings that worked well in this study. Further research is needed to determine ideal embeddings that capture the jargon and biased language within specific contexts, such as clinical notes or legal briefs.

For now, we suggest using monolingual embedding models with some validation and cleaning of input data for embedders, as multilingual models without any curation/validation steps did not give good results. It is possible that better multilingual models will exist in the future, and multilingual and multicultural bias applications will be more feasible at that point in time.

V. Conclusion

In our study, we aimed to understand the efficacy of pre-trained transformer models in detecting various biases in textual data, specifically from the RedditBias database. Our findings indicate that BERT embeddings, particularly the mini BERT embedding, outperformed RoBERTa embeddings in classifying bias types. The performance was especially notable with certain choices of the 'k' parameter in KNN models, with k=3 yielding the best results. It was observed

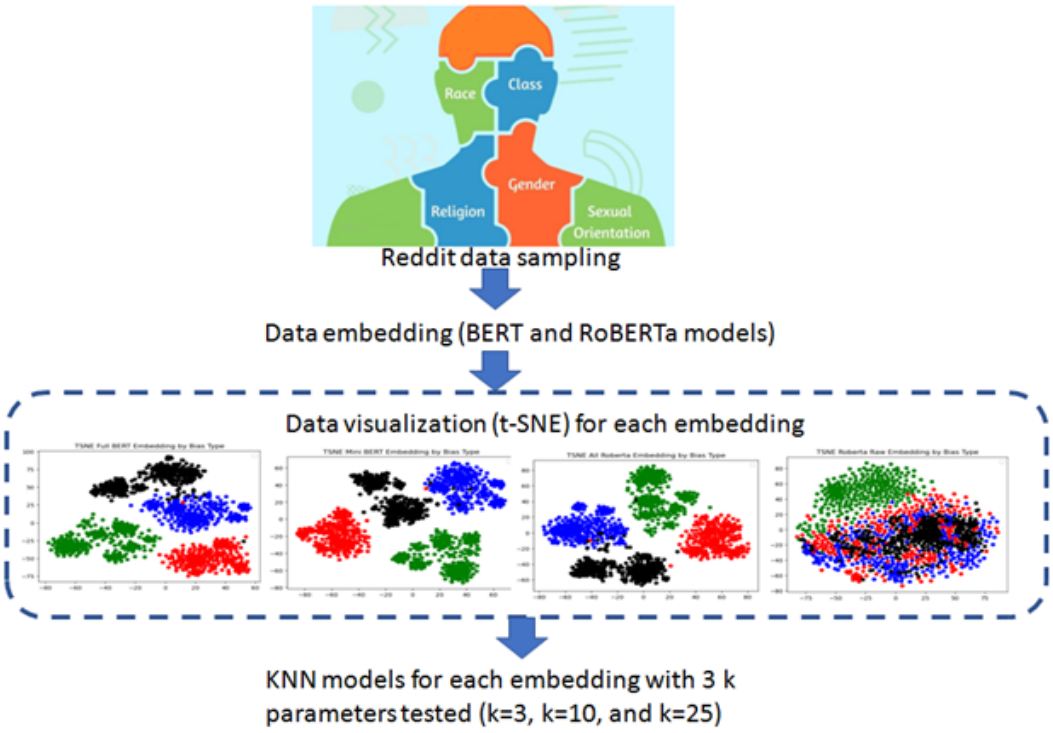

Fig. 1: Flow of data processing from sampling to embedding with BERT models to visualizing embeddings to creating KNN models with varying k parameter.

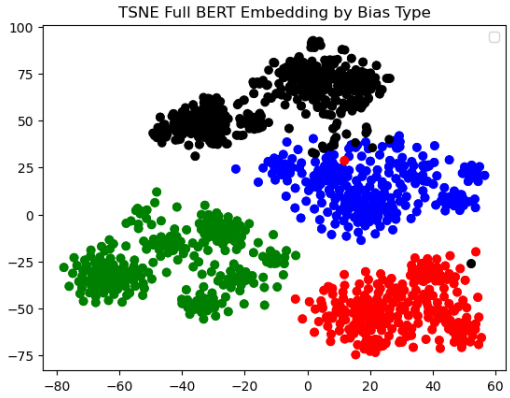

Fig. 2: t-SNE embedding of all-mpnet-base-v2 model of our Reddit sample.

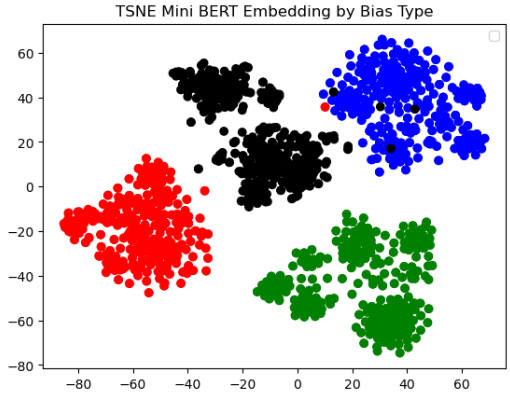

Fig. 3: t-SNE embedding of all-MiniLM-L6-v2 model of our Reddit sample.

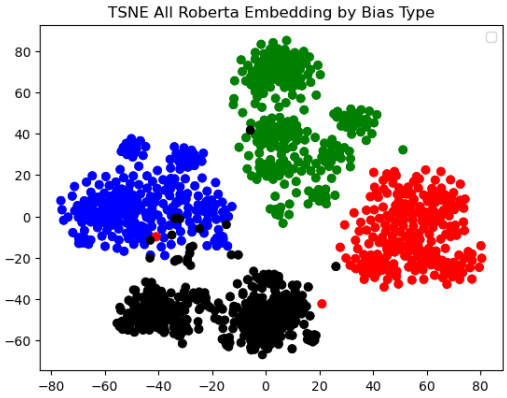

Fig. 4: t-SNE embedding of all-roberta-large-v1 model of our Reddit sample.

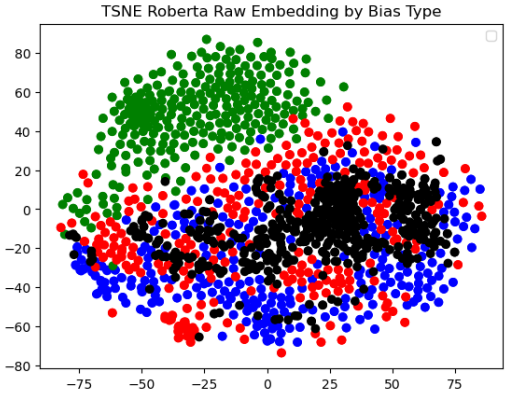

Fig. 5: t-SNE embedding of xlm-roberta-base model of our Reddit sample.

that the raw RoBERTa model, trained on a large multilingual dataset, was less effective compared to other embeddings, emphasizing the potential importance of data curation and language specificity in training models for bias detection.

The results highlight the robustness of the BERT models and the single-language RoBERTa model in handling real-world English text nuances, such as misspellings and grammatical errors. However, multilingual models that lack a validation or cleaning phase in their training data exhibited reduced effectiveness. This suggests that while there's potential for future multilingual models to improve, the current recommendation would be to employ monolingual embedding models with validated and curated input data.

Given the success of the models on Reddit data, there's optimism about applying the methodology to field-specific texts, such as educational, medical, and legal notes. However, careful consideration is required due to potential challenges posed by field-specific jargon and subtler forms of bias. Regardless, the methodology offers the potential for institutions to identify and rectify biases, ultimately improving outcomes for the diverse communities they serve.

Looking forward, as technologies and embeddings evolve, future research should explore optimized embeddings to capture nuanced and context-specific biased language.

VI. Compliance with Ethical Standards

Ethical and informed consent for data used: Not Applicable

VII. Funding

No funding

VIII. Competing Interests

The authors declare that they have no conflict of interest. Data availability and access: The datasets created and/or analyzed during the current investigation are available upon reasonable request from the corresponding author.

Acknowledgment

This work was supported by the Laboratory of Energetic System Modelling (LMSE) of the University of Biskra, Algeria, under the patronage of the General Directorate of Scientific Research and Technological Development (DGRSDT) in Algeria. The research project was approved by the Ministry of Higher Education and Scientific Research in Algeria, under the number A01L08UN070120220003.

References

[1] Y. Himeur, A. Alsalemi, A. Al-Kababji, F. Bensaali, A. Amira, C. Sardianos, G. Dimitrakopoulos, and I. Varlamis, "A survey of recommender systems for energy efficiency in buildings: Principles, challenges and prospects," *Information Fusion*, vol. 72, pp. 1–21, 2021.

[2] C. Sardianos, I. Varlamis, C. Chronis, G. Dimitrakopoulos, A. Alsalemi, Y. Himeur, F. Bensaali, and A. Amira, "The emergence of explainability of intelligent systems: Delivering explainable and personalized recommendations for energy efficiency," *International Journal of Intelligent Systems*, vol. 36, no. 2, pp. 656–680, 2021.

[3] Y. Singh, C. Farrelly, Q. A. Hathaway, A. Choudhary, G. Carlsson, B. Erickson, and T. Leiner, "The role of geometry in convolutional neural networks for medical imaging," *Mayo Clinic Proceedings: Digital Health*, vol. 1, no. 4, pp. 519–526, 2023.

[4] H. H. D. M. S. S. S. Y. H. M. A. A. Faiza Farhat, Emmanuel Sirimal Silva and A. Zafar, "Analyzing the scholarly footprint of chatgpt: Mapping the progress and identifying future trends," *Frontiers in Artificial Intelligence, section Natural Language Processing*, pp. 1–22, 2023.

[5] Y. Singh, C. M. Farrelly, Q. A. Hathaway, T. Leiner, J. Jagtap, G. E. Carlsson, and B. J. Erickson, "Topological data analysis in medical imaging: current state of the art," *Insights into Imaging*, vol. 14, no. 1, pp. 1–10, 2023.

[6] Y. Himeur, S. S. Sohail, F. Bensaali, A. Amira, and M. Alazab, "Latest trends of security and privacy in recommender systems: a comprehensive review and future perspectives," *Computers & Security*, vol. 118, p. 102746, 2022.

[7] C. Sardianos, I. Varlamis, G. Dimitrakopoulos, D. Anagnostopoulos, A. Alsalemi, F. Bensaali, Y. Himeur, and A. Amira, "Rehab-c: Recommendations for energy habits change," *Future Generation Computer Systems*, vol. 112, pp. 394–407, 2020.

[8] D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman, "Nuanced metrics for measuring unintended bias with real data for text classification," in *Companion proceedings of the 2019 world wide web conference*, 2019, pp. 491–500.

[9] E. Goffman. Simon and Schuster, 2009.

[10] J. Y. Edgoose, M. Quiogue, and K. Sidhar, *Family practice management*, vol. 26, no. 4, pp. 29–33, 2019.

[11] J. A. Clarke, *Nw. UL Rev.*, vol. 113, p. 505, 2018.

[12] L. Aron and P. Loprest, *The future of Children*, pp. 97–122, 2012.

[13] M. L. Hatzenbuehler, *American Psychologist*, vol. 71, no. 8, p. 742, 2016.

[14] L. C. Van Boekel, E. P. Brouwers, J. Van Weeghel, and H. F. Garretsen, *Drug and alcohol dependence*, vol. 131, no. 1-2, pp. 23–35, 2013.

[15] R. D. Ashford, A. M. Brown, J. McDaniel, and B. Curtis, *Substance use & misuse*, vol. 54, no. 8, pp. 1376–1384, 2019.

[16] I. Kim, T. S. Field, D. Wan, K. Humphries, and T. Sedlak, *Canadian Journal of Cardiology*, 2022.

[17] J. C. McSweeney, A. G. Rosenfeld, W. M. Abel, L. T. Braun, L. E. Burke, S. L. Daugherty, ..., and J. F. Reckelhoff, *Circulation*, vol. 133, no. 13, pp. 1302–1331, 2016.

[18] N. K. Wenger, L. Speroff, and B. Packard, *New England Journal of Medicine*, vol. 329, no. 4, pp. 247–256, 1993.

[19] M. Sun, T. Oliwa, M. E. Peek, and E. L. Tung, "Negative patient descriptors: Documenting racial bias in the electronic health record: Study examines racial bias in the patient descriptors used in the electronic health record," *Health Affairs*, vol. 41, no. 2, pp. 203–211, 2022.

[20] M. C. Beach *et al.*, "Testimonial injustice: linguistic bias in the medical records of black patients and women," *Journal of general internal medicine*, vol. 36, no. 6, pp. 1708–1714, 2021.

[21] A. P. Goddu *et al.*, "Do words matter? stigmatizing language and the transmission of bias in the medical record," *Journal of general internal medicine*, vol. 33, no. 5, pp. 685–691, 2018.

[22] S. M. Carroll, "Respecting and empowering vulnerable populations: contemporary terminology," *The Journal for Nurse Practitioners*, vol. 15, no. 3, pp. 228–231, 2019.

[23] A. C. Gill *et al.*, "Longitudinal outcomes one year following implicit bias training in medical students," *Medical Teacher*, pp. 1–8, 2022.

[24] M. Ruben and N. S. Saks, "Addressing implicit bias in first-year medical students: a longitudinal, multidisciplinary training program," *Medical Science Educator*, vol. 30, no. 4, pp. 1419–1426, 2020.

[25] S. M. Phelan *et al.*, "Medical school factors associated with changes in implicit and explicit bias against gay and lesbian people among 3492 graduating medical students," *Journal of general internal medicine*, vol. 32, no. 11, pp. 1193–1201, 2017.

[26] S. S. Sohail, F. Farhat, Y. Himeur, M. Nadeem, D. Ø. Madsen, Y. Singh, S. Atalla, and W. Mansoor, "The future of gpt: A taxonomy of existing chatgpt research, current challenges, and possible future directions," *Current Challenges, and Possible Future Directions (April 8, 2023)*, 2023.

[27] S. S. Sohail, D. Ø. Madsen, Y. Himeur, and M. Ashraf, "Using chatgpt to navigate ambivalent and contradictory research findings on artificial intelligence," *Available at SSRN 4413913*, 2023.

[28] S. S. Sohail, F. Farhat, Y. Himeur, M. Nadeem, D. Ø. Madsen, Y. Singh, S. Atalla, and W. Mansoor, "Decoding chatgpt: A taxonomy of existing research, current challenges, and possible future directions," *Journal of King Saud University-Computer and Information Sciences*, p. 101675, 2023.

[29] N. D. Y. H. Feriel Khennouche, Youssef Elmir and A. Amira.

[30] J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, no. 1, 2003, pp. 29–48.

[31] J. Devlin *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[32] F. Li *et al.*, "Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: an empirical study," *JMIR medical informatics*, vol. 7, no. 3, p. e14830, 2019.
[33] M. Mozafari *et al.*, "A bert-based transfer learning approach for hate speech detection in online social media," in *International Conference on Complex Networks and Their Applications*. Springer, 2019, pp. 928–940.
[34] T. Bolukbasi *et al.*, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *Advances in neural information processing systems*, vol. 29, 2016.
[35] K. Kowsari *et al.*, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.
[36] Z. Ahmed *et al.*, "Tackling racial bias in automated online hate detection: Towards fair and accurate detection of hateful users with geometric deep learning," *EPJ Data Science*, vol. 11, no. 1, p. 8, 2022.
[37] N. Janasik, T. Honkela, and H. Bruun, "Text mining in qualitative research: Application of an unsupervised learning method," *Organizational Research Methods*, vol. 12, no. 3, pp. 436–460, 2009.
[38] H. Kheddar, Y. Himeur, S. Al-Maadeed, A. Amira, and F. Bensaali, "Deep transfer learning for automatic speech recognition: Towards better generalization," *arXiv preprint arXiv:2304.14535*, 2023.
[39] M. Yousefi-Azar and L. Hamey, "Text summarization using unsupervised deep learning," *Expert Systems with Applications*, vol. 68, pp. 93–105, 2017.
[40] Z. Chen *et al.*, "The lao text classification method based on knn," *Procedia Computer Science*, vol. 166, pp. 523–528, 2020.
[41] G. Guo *et al.*, "Using k nn model for automatic text categorization," *Soft Computing*, vol. 10, pp. 423–430, 2006.
[42] L. S. Larkey and W. B. Croft, "Combining classifiers in text categorization," in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, 1996, pp. 289–297.
[43] K. Shah *et al.*, "A comparative analysis of logistic regression, random forest and knn models for the text classification," *Augmented Human Research*, vol. 5, pp. 1–16, 2020.
[44] K. Almgren, M. Kim, and J. Lee, "Mining social media data using topological data analysis," in *2017 IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE, 2017, pp. 144–153.
[45] P. Doshi and W. Zadrozny, "Movie genre detection using topological data analysis," in *International Conference on Statistical Language and Speech Processing*. Springer, Cham, 2018, pp. 117–128.
[46] H. Heuer, "Text comparison using word vector representations and dimensionality reduction," *arXiv preprint arXiv:1607.00534*, 2016.
[47] R. González-Márquez, P. Berens, and D. Kobak, "Two-dimensional visualization of large document libraries using t-sne," in *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*, 2022.
[48] S. Barikeri, A. Lauscher, I. Vulić, and G. Glavaš, "Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models," *arXiv preprint arXiv:2106.03521*, 2021.
[49] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.
[50] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. 11, 2008.
[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.
[52] K. Fukunaga and P. M. Narendra, "A branch and bound algorithm for computing k-nearest neighbors," *IEEE transactions on computers*, vol. 100, no. 7, pp. 750–753, 1975.
[53] H. A. Abu Alfeilat, A. B. Hassanat, O. Lasassmeh, A. S. Tarawneh, M. B. Alhasanat, H. S. Eyal Salman, and V. S. Prasath, "Effects of distance measure choice on k-nearest neighbor classifier performance: a review," *Big data*, vol. 7, no. 4, pp. 221–248, 2019.
[54] G. G. Enas and S. C. Choi, *Choice of the smoothing parameter and efficiency of k-nearest neighbor classification*. Pergamon, 1986, pp. 235–244.