DOI: 10.1590/s0103-4014.202438111.019

Mitigação de viés de *datasets* multimodais em um classificador de categorias urbano-sociais

Luciano C. Lugli [I]
Daniel Abujabra Merege [II]
Rafael Pillon Almeida [III]

Introdução

Piaget (1974; 1983), referência na psicogênese pedagógica, dedicou suas pesquisas à construção do conhecimento, denominada por ele *epistemologia genética*, sendo "[...] a descoberta das várias variáveis de conhecimento, desde seu modo elementar, e seguindo sua evolução epistêmica do nascimento aos níveis seguintes, edificando singularmente a ética no ser humano" (Piaget, 1983, p.3). Segundo Piaget (1991, p.23), o desenvolvimento da moralidade é vasto, e assevera que "[...] a moral é dotada de um sistema de regras, e a essência ética da moralidade deve ser procurada no respeito que o indivíduo adquire por essas regras".

Na perspectiva do citado referencial sobre a moralidade e a construção do conhecimento por Piaget, e direcionando os objetivos desse projeto na mitigação de viés (*bias*) sobre *datasets* multimodais (Yoon et al.,, 2017; Jiang et al., 2019; Testuggine et al., 2020) a um classificador de categorias urbanas, escutar e entender as várias manifestações digitais de pessoas sobre as problemáticas municipais, demanda uma interpretação implicativa do (inter)locutor sobre tal tema, na identificação dos itens subjetivos do sujeito no discurso que, pautada pela parametrização linguística na teoria do letramento (Bakhtin, 2011; 2018; Pêcheux, 1988; 1990), sejam críveis e conclusivas sobre a convergência contextual das categorias nos textos inferidos.

Contudo, a correlação entre estratificar dados de treinamento e entender dados de testes deve ser heterogênea quanto à subjetividade individual das manifestações textuais, e homogênea quanto à sujeição de convergência coletiva aos paradigmas sociais que especificam o caráter na interpretação cognitivo-humana sobre essas manifestações textuais. Destarte, convergir coletivamente aos valores morais e sociais em Aprendizado de Máquina – AM (*Machine Learning* – ML, em inglês), através de uma abordagem analítica em Processamento de Linguagem Natural – PLN (*Natural Language Processing* – NLP, em inglês), demanda uma identidade racional aos preceitos político-públicos e sociomorais

que desenvolvedores de software conflitam e consonam sobre um viés mínimo e permissível a pluralidade de dados seja em treinamento, seja em teste (Mellet et al., 2014; O'Neil, 2016; Noble, 2018; Ma, 2018).

Mitigar viés, no caso desse projeto, é uma tarefa que, mesmo inicialmente dispendiosa, é necessária quando se pretende inferir detalhadamente em categorias que têm uma tênue significância em sua semântica (*e.g.*: verificar se um texto é mais/menos *político*, mas tem características de *governança*, e ênfase a fatores financeiros da *economia*; ou identificar contextos compelidos a textos sobre *educação*, mas notoriamente relativo à participação e promoção da *cultura*).

Objetivos

Este projeto objetiva relacionar os pressupostos psicogenéticos do desenvolvimento sociomoral e da linguística discursiva na mitigação do viés (*bias*) em dados de aprendizado, avaliação e testes amostrais, sobre a interferência direta no desenvolvimento de um interpretador de categorias urbanas, baseado no mecanismo de atenção adaptado *Transformers* (Vaswani et al., 2017) – uma arquitetura que infere direta e indiretamente sobre o processamento de dados paralelos distribuídos em *tokens*, e que são contextualizados pela convergência mínima relacional. Trata-se de verificar a ética existente durante o crivo e critério em que *datasets* multimodais suportam e sustentam a classificação errônea de categorias, assim prejudicando a parametrização da precisão final na interpretação.

Referenciais teóricos

Entender e elucidar as várias manifestações públicas das problemáticas dentro de uma cidade/centro urbano demanda uma caracterização inerente ao (inter) locutor sobre tal tema, numa tentativa em identificar os itens subjetivos do sujeito no discurso – *quem fala, como fala, o que fala, de onde fala, para quem fala...*) (Bakhtin, 2011). Necessita-se de um apoio acadêmico de outras áreas do conhecimento que, se normalizados em suas respectivas naturezas científicas contextuais, tornam-se integrais ao mesmo tempo, *técnica-tecnológica* e *linguística-social*.

Desse modo, são descritas nesta seção as referências teóricas que fundamentam a base linguística moral que vieses são construídos a partir de um contexto cultural, social e técnico. Quando tratamos de uma abordagem analítica textual em PLN voltada a textos manifestados diretamente por pessoas em plataformas virtuais de interação pública, o contexto compreendido deve considerar uma interpretação que convalida de modo correlacional os enunciados estruturados esquematicamente pelo léxico, vocabulário e significação semântica (Brait, 2005), permitindo à classificação em AM identificar os pormenores de uma categoria a outra.

Mitigar o viés nessa perspectiva convalida a convergência em que tanto a compreensão contextual de cada responsável pela discriminação de classes nas bases de dados quanto o critério que se estabelece no entendimento das classes urbano-sociais são inerentemente relacionadas à formação integral (*e.g.*: moral) cidadã (Angwin et al., 2017; Assimakopoulos et al., 2020; Blodgett et al., 2020).

Das premissas de Piaget (1994, p.23), o desenvolvimento da moralidade é bem amplo, não se refere a apenas obedecer determinadas regras, mas sim compreender a razão pela qual o sujeito a obedece; portanto, ele assegura que "[...] toda moral consiste num sistema de regras, e a essência de toda moralidade deve ser procurada no respeito que o indivíduo adquire por essas regras".

Nessa perspectiva, suas pesquisas (Piaget, 1994) mostram que o processo de construção da moralidade é uma tarefa ardilosa, especialmente no contexto social em que as prerrogativas do desenvolvimento e sociointerativo cognitivo se sobrepõem ao desenvolvimento moral. Mesmo tendo referenciais nos documentos oficiais que fomentam e fundamentam uma metodologia pedagógica na construção moral desde a infância na educação infantil, as relações interpessoais e o desenvolvimento integral do ser humano são prejudicados por condutas e comportamentos antissociais emergidos pelo viés.

Assim, uma análise discursiva se mostra necessária no modo anunciativo que as manifestações digitais têm se perpetrado nos vários textos publicados nas redes sociais, tanto pela autoria da defesa do discurso quanto pelo ente moral que aquilo se refere. E nesses pressupostos, tal análise se relaciona diretamente à uma relação de categorias urbano-sociais que tratam sobre várias temáticas entre cidade-cidadão, como questões relacionadas sobre a governabilidade e gerenciamento dos serviços empregados em um município.

Para verificar e denotar as categorias urbanas envoltas ao contexto compreendido em cada texto, é necessário agregar a semântica da subjetividade linguística (Lahire, 1990; 1993a) com os parâmetros intrínsecos do *mecanismo de atenção adaptada* da arquitetura *Transformers*, que discretiza a significância e sentidos das palavras (*n*-gramas) através de um modelo denominado *integração relacional paralela de construção linguística*.

Entendendo e elucidando a linguagem nas redes sociais por Bakhtin

A língua só existe em razão do uso que locutores (quem fala ou escreve) e interlocutores (quem lê ou escuta) fazem dela em situações (prosaicas ou formais) de comunicação (manifestações digitais). O ensinar, o aprender e o empregar a linguagem passa necessariamente pelo sujeito, o agente das relações sociais e o responsável pela composição e pelo estilo dos discursos. Esse sujeito se vale do conhecimento de enunciados anteriores para formular suas falas e redigir seus textos. Além disso, um enunciado sempre é modulado pelo falante para o contexto social e cultural, caso contrário, ele não será compreendido (uma referência da descaracterização público-social para a categoria "*Não urbano*") (Bakhtin, 2011; 2018; Brait, 2009; Brandist; Tihanov, 2000).

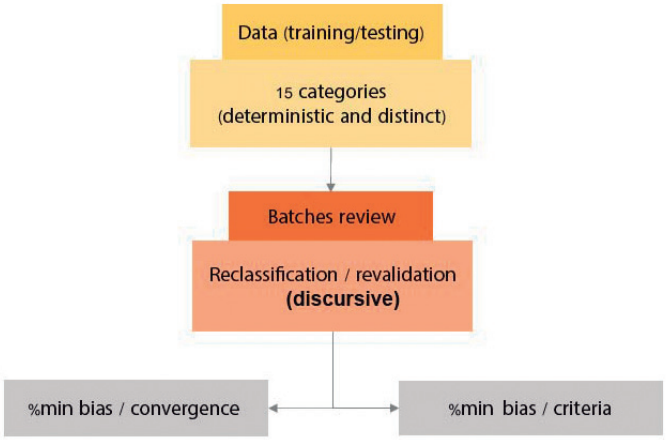
Nessa relação dialógica entre locutor e interlocutor no meio social, em que o verbal e o não verbal influenciam de maneira determinante a construção dos enunciados, a interação por meio da linguagem se dá num contexto em que todos participam em condição de igualdade (homogeneidade do direito à manifestação). Aquele que enuncia seleciona palavras apropriadas para formular uma

mensagem compreensível para seus destinatários (*i.e.*: comentar diretamente sobre uma categoria urbana). Por outro lado, o interlocutor interpreta e parametriza o significado do enunciado (por meio de sua inferência sistêmica – *e.g.*: PLN).

Métodos

A metodologia planejada e projetada para a mitigação de viés se estrutura nas bases de dados de treinamento e teste em um modelo referencial de AM para a classificação de categorias multimodais.

Para contextualizar as características da problemática envolta ao viés (*bias*) em PLN, implementou-se um classificador de categorias urbano-socias, baseado no mecanismo de atenção adaptada Transformers (Vaswani et al., 2017), que é parametrizada segundo características da classificação multiclasses sobre o set de bibliotecas *simpletransformers* à biblioteca nativa *transformers* e bases na língua portuguesa ao *HuggingFace/BERTimbau* para pré-treino e tokenizer (*neuralmind/bert-base-portuguese-cased*) (Souza et al., 2020), e na língua espanhola ao *HuggingFace/BETO* para pré-treino e tokenizer (*dccuchile/bert-base-spanish-wwm-cased*) (Cañete et al. 2020), ambos modelos pré-treinados do *HuggingFace/BERT* (Devlin et al., 2018).



Fonte: Autoria própria.

Figura 1 – Diagrama sequencial da estratégia de mitigação de viés sobre as bases de dados de treino e teste.

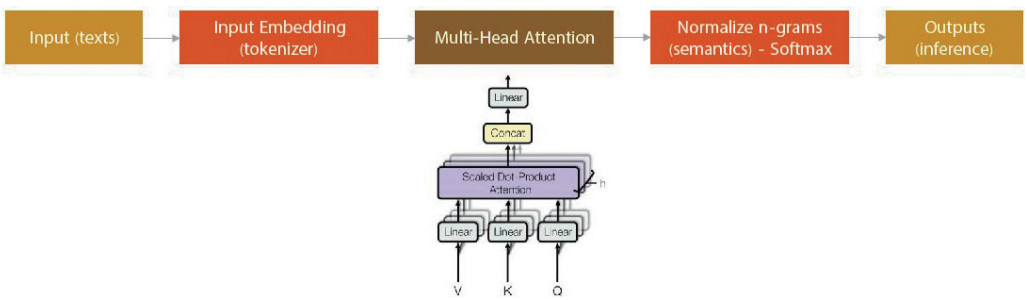
O classificador de categorias urbano-sociais é usado como referência de classificação multiclasses, uma vez que reflete o caráter e contexto público-social em um município, em que para ser categorizado corretamente, o texto precisa mostrar a adequação e associação a um fator/função urbana-social, e dirimir alguns questionamentos, como: Este texto se relaciona a algum aspecto da gestão pública? Quais repercussões têm o conteúdo contextual deste texto na gestão e nos serviços públicos? Este texto ajuda no aprimoramento de serviços públicos e sociais? Este texto provê uma gestão participativa na cidade (dos cidadãos) socialmente?

E nesse classificador, são relacionadas 15 categorias urbano-sociais, definidas e descritas a um fator/função social-urbana ("*Cultura e recreação*", "*Economia*", "*Educação*", "*Gestão de resíduos*", "*Governança*", "*Habitação/Assistência social*", "*Infraestrutura urbana*", "*Meio-ambiente*", "*Mobilidade*", "*Política*", "*Saúde*", "*Saneamento básico*", "*Segurança*", "*Turismo*" e, quando um texto descaracteriza qualquer uma das categorias *a priori*, sem coesão e coerência urbano-social, ausente de argumentos classificatórios às categorias, é definido como "*Não urbano*"). Um texto é classificado, de acordo com seu *corpus* (conteúdo e contexto), e semântica temática (expressões que relacionam direta e indiretamente o *corpus*) em um classe (com % de confidência máximo para uma categoria) (*e.g.*: um texto é classificado somente como "Educação"), ou em mais classes (com % de confidência mínimo para mais de uma categoria) (*e.g.*: um texto é classificado parcialmente como "Educação" e "Política", por prover características linguísticas/lexicais destas duas categorias).

Mecanismo de atenção adaptada para classificação de textos – Transformers

Transformers é uma arquitetura de redes neurais que prioriza o paralelismo da "*atenção*", ao invés da recorrência, o que no contexto *seq2seq*, observou-se que *layers* de atenção melhoram muito o desempenho das redes (Yoon et al., 2017; Basu, 2020) em paralelizar o léxico em PLN.

Tal paralelização permite inferir na análise textual de modo correlacional, o que significa que os *n*-gramas (*e.g.*: palavras, predicados, orações) são analisados no esquema estrutural lexical em: *"um a um"* e *"um para um"*, inferindo dessa forma em funções que caracterizam se um texto tem mais probabilidade de ser classificado conforme pretendido *a priori* (Jian et al., 2019; Kielay et al., 2020).



Fonte: Autoria própria.

Figura 2 – Diagrama geral da arquitetura adaptada baseada em Transformers.

A priori, tal interpretador resultava em predições que tinham uma conotação correlacional de critério erroneamente enviesado, entravando relativamente os percentuais de predições de categorias às bases de treinamento (*i.e.*: 1ª cate-

goria errada e 2ª categoria exata; ou conotação contextual direcionada mais a uma categoria e menos a outra).

Mitigação de viés – revisão, revalidação e reclassificação

Para as 15 categorias, os *batches* (terminologia usada para descrever conjuntos de dados) eram iniciados com uma quantidade crescente proporcional balanceada de textos/categoria, isto é, inicialmente um *batch* com a mesma quantidade relacional de textos por categoria (balanceados), e crescente nos demais *batches* devido ao desbalanceamento natural em que um texto era reclassificado dado o critério de mitigação de viés.

Esquematicamente sobre a estrutura de revisão de dados, as bases de treinamento se dividiram inicialmente em bases parciais de revisão/reclassificação em *batches* de quantidade definida semanalmente. Cada base parcial semanal subsequente tinha uma distribuição de textos/categoria que dependia da revisão da base atual, provendo assim um balanceamento contínuo de carga para manter a quantidade homogênea. Quanto a qualidade, buscou-se relacionar textos de compreensão contextual agregada a semântica da subjetividade linguística (Lahire, 1990; Gillani; Levi, 2019; Hanna et al., 2020; Leins et al., 2020) com os parâmetros intrínsecos do *mecanismo de atenção adaptada*, que discretiza a significância e sentidos das palavras (*n*-gramas). Isso significa que os textos deveriam ter parágrafos, predicados ou orações que, para além da semântica, apresentassem um léxico de *tags* relacionados à sua respectiva categoria.

Três times de revisão foram formados para a tarefa de reclassificação dos textos em cada *batch* semanal: '*T1*' era um time diversificado com especialistas de IA/AM e PLN; '*T2*' era um time diversificado com especialistas em linguística (licenciados em Letras); '*T3*' era um time heterogêneo representando minorias de classe/comunidade social.

As revisões eram estruturadas entre *convergência plena* (concordância integral dos membros de revisão), *convergência parcial* (concordância majoritária dos membros de revisão) e *divergência plena* (discordância integral dos membros de revisão). Nesta última, uma revisão de pormenores é realizada texto a texto para identificar os vários vieses atribuídos e discutir sob a abordagem analítica do discurso linguístico (Lahire, 1993b; 1993c), e no referencial da psicogênese epistêmica sobre a ética, quais argumentos eram defendidos às suas respectivas classificações, concluindo com isso, um critério de assentimento de categoria final pelos membros.

Buscou-se nesse procedimento de revisão dos *batches*, discriminar as variedades de vieses que desenvolvedores precisam evitar: viés de algoritmo, viés de amostra, viés de (pré)conceito e viés de medida (Brown et al., 2019; Field et al., 2021a; 2021b).

Viés de algoritmo: refere-se à própria propriedade do algoritmo sobre a tendência ao desempenho do modelo entre treinamento e teste. Deve-se então, encontrar o equilíbrio entre este viés e a variância de dados (balancear a quantidade e qualidade de corpus);

Viés de amostra: refere-se à qualidade representativa da precisão amostral em que os dados são treinados. Deve-se incorrer a amostras de populações heterogêneas na qualidade, para validar sua representatividade homogênea sobre o tamanho amostral;

Viés de (pré)conceito: refere-se ao conteúdo ético sobre a relação "*comportamento e cognição*" que desenvolvedores têm em suas conceituações e concepções político-públicas e sociais. Mitigar este viés requer restrições atribuídas às maneiras sociomorais que os dados são classificados;

Viés de medida: refere-se à medida/métrica que o dado é desempenhado, com resultados sistematicamente distorcidos. Deve-se equilibrar o modo técnico entre a aquisição e análise de dados para padronizar naturezas distintas de dados.

Baseado nessa metodologia, ressalta-se a relevância sistêmica que os dados são tratados para a tarefa de classificação de categorias urbano-sociais, a respeito da Estratégia Brasileia de IA (Brasil/EBIA, 2021). Como descrito em seu capítulo "Legislação, Regulação e Uso Ético", é necessário identificar vieses envolvidos em interpretações decisórias (reconhecimento de padrões) e desafiar tais interpretações, quando cabível ao caráter ético em ML, pautado por discussões sobre sistemas de IA confiáveis e centradas na relação humano-máquina (*trustworthy H-M AI*).

Resultados

Os resultados de revisão/reclassificação dos dados de treinamento à referida arquitetura proposta mostram uma acurácia assertiva de ~90,0%, *post* mitigação do viés, considerando a convergência contextual cruzada, dada a discriminação de dados por inferência paralela de critério. Isso identifica uma progressão dos percentuais de revalidação dos *batches*, discriminados em *convergência plena* (*CPe*), *convergência parcial* (*CPa*) e *divergência plena* (*DPe*).

Ainda, a validação da regularização dos dados modelados foi feita baseando-se na razão entre viés e variância. Um viés baixo significa que o modelo aprende, mas de um modo ajustado na relação entre as variáveis e as predições das bases de dados de teste, o que inviabiliza a discriminação de dados distintos dos associados, *a priori*, na base de treino, desenvolvendo-se uma variância (Blodgett et al., 2021; Castelle, 2018; Hutchinson e al., 2020; Jiang; Fellbaum, 2020; Lepori, 2020; Motha, 2020).

Logo, para manter a razão regulatória no modelo, diminui-se a variância pela heterogeneidade de dados relacionais de cada viés (cada time de revisão), conseguindo convergir e generalizar a dados de testes, *a posteriori*, por uma *ridge regression* (Hilt; Seegrist, 1977) parametrizada nos dados revisados.

Para as duas línguas analisadas, portuguesa e espanhola, foram fixados 7 e 5 *batches* de treino/teste, respectivamente, uma vez que cada base apresentava uma associação distinta de quantidade geral, quantidade por classe, *tokenizers* sobre as bases pretendidas, e estratégias de revisão dedicadas a cada demanda.

Para ambas as Tabelas 1 e 2, a discriminação de revisão para os percentuais de convergência e de critério são definidas pelos três times de revisão 'T1', 'T2' e 'T3'. Quando os três times se equiparam, tem-se a *convergência plena (CPe)*; quando algum dos times não se equipara aos demais, tem-se a *convergência parcial (CPa)*, em que são revisadas três situações de vieses distintos; e quando nenhum dos times se equipara, tem-se a *divergência plena (DPe)*. A linha 01 mostra a convergência plena – *CPe* (os times T1, T2, T3 têm os mesmos vieses); A linha 02 mostra a divergência plena – *DPe* (os times T1, T2, T3 têm vieses diferentes um do outro); As linhas 03–05 mostram a convergência parcial – *CPa* (as três situações em que dois times têm o mesmo viés, porém difere do outro).

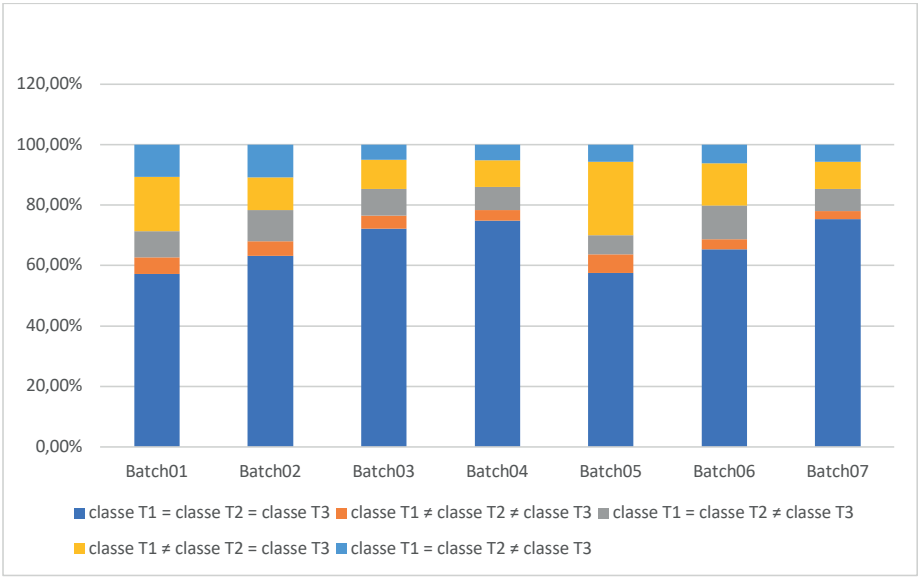
Na Tabela 1 verificam-se os resultados percentuais entre os *batches* 01–07 e as cinco correspondências de convergência/critério de revalidação para a língua portuguesa.

Tabela 1 – Distribuição dos batches e respectivos percentuais para cada discriminação definida pela estratégia de revisão (língua portuguesa)

Critério de comparação	*Batch*01	*Batch*02	*Batch*03	*Batch*04	*Batch*05	*Batch*06	*Batch*07
classe T1=classe T2=classe T3	57,20%	63,20%	72,20%	74,80%	57,60%	65,40%	75,40%
classe T1≠classe T2≠classe T3	5,60%	4,80%	4,40%	3,60%	6,20%	3,40%	2,60%
classe T1=classe T2≠classe T3	8,60%	10,40%	8,80%	7,60%	6,20%	11,00%	7,40%
classe T1≠classe T2=classe T3	18,00%	10,80%	9,60%	8,80%	24,40%	14,00%	9,00%
classe T1=classe T2≠classe T3	10,60%	10,80%	5,00%	5,20%	5,60%	6,20%	5,60%

Fonte: Autoria própria.

Percebe-se que, na língua portuguesa, os dados da Tabela 1/Figura 3 mostram um progresso no percentual para o caso de convergência plena (CPe) entre os *batches* 01 e 04 (de 57,20% a 74,80%). O *batch* 05 apesentava dados distintos do viés treinado anteriormente, promovendo uma heterogeneidade na reclassificação dos dados (variância > viés), e o critério de comparação entre os times T1, T2 e T3 se restabeleceu, e evoluiu normalmente conforme mais dados eram revisados entre os *batches* 05 e 07. Ao final da revisão dos 7 *batches*, a CPe obteve um percentual de 75,40%. Para o critério de comparação da divergência plena (DPe), o percentual entre os *batches* 01 e 07 diminuiu significativamente, e finalizou em 2,60% de mínima. Mesmo na mesma situação da razão regulatória (viés e variância) a partir do *batch* 05, este percentual é bem abaixo da revisão estimada, não chegando aos 7% de máxima, e denota que os vieses dos três times são concisos e convergentes na categorização.



Fonte: Autoria própria.

Figura 3 – Gráfico da distribuição dos batches e respectivos percentuais na língua portuguesa.

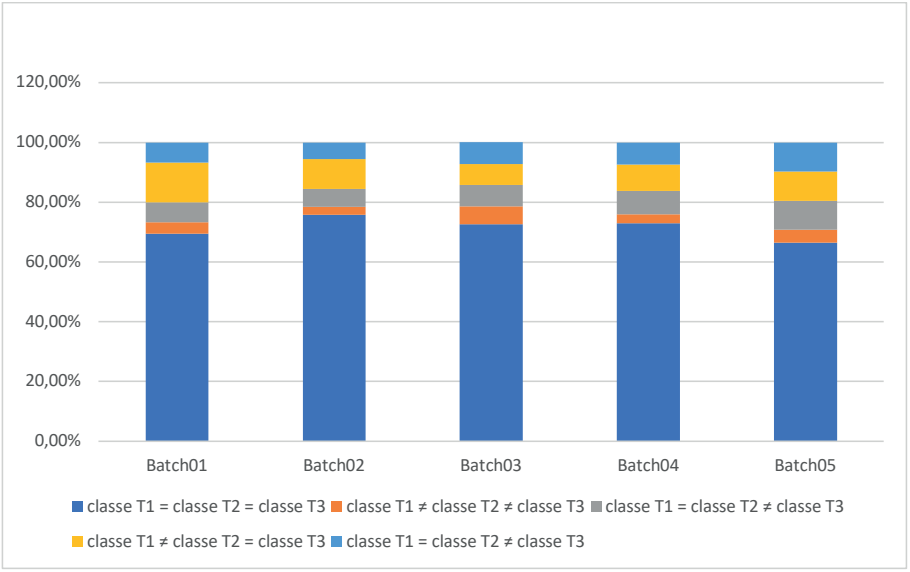
Finalmente, para o caso do critério de comparação com convergência parcial (CPa), dois *batches* tiveram resultados previsíveis pela razão regulatória desbalanceada (viés > variância), o *batch* 01, por se tratar da revisão inicial, e o *batch* 05, no balanceamento da razão regulatória (viés < variância). Para as demais revisões dos *batches*, os percentuais de convergência parcial entre os três times se mantiveram numa proporção plausível e permissível.

Na Tabela 2 verificam-se os resultados percentuais entre os *batches* 01–05 e as cinco correspondências de convergência/critério de revalidação para a língua espanhola.

Tabela 2 – Distribuição dos batches e respectivos percentuais para cada discriminação definida pela estratégia de revisão (língua espanhola).Fonte: autoria própria.

Critério de comparação	Batch*01*	Batch*02*	Batch*03*	Batch*04*	Batch*05*
classe T1=classe T2=classe T3	69,50%	75,80%	72,70%	73,00%	66,40%
classe T1≠classe T2≠classe T3	3,80%	2,60%	5,90%	3,00%	4,40%
classe T1=classe T2≠classe T3	6,70%	6,00%	7,10%	7,80%	9,60%
classe T1≠classe T2=classe T3	13,30%	10,00%	7,10%	8,80%	9,80%
classe T1=classe T2≠classe T3	6,70%	5,60%	7,30%	7,40%	9,80%

Fonte: Autoria própria.



Fonte: Autoria própria.

Figura 4 – Gráfico da distribuição dos batches e respectivos percentuais na língua espanhola.

Na língua espanhola, os dados da Tabela 2/Figura 4 mostram progressos diferentes dos da língua portuguesa. Inicialmente, foram fixados somente 5 *batches*, e o primeiro (*batch* 01) era significativamente menor que os demais *batches*, por se tratar de um teste de validação com categorias mais críticas (*e.g.*: muitas classes erroneamente em *'Not Urban'*). Consequentemente, as demais revisões foram fomentadas na caracterização do *batch* 01, e no critério de comparação de convergência plena (CPe), o percentual variou entre 66,40% e 75,80%, e o último *batch* revisado apresentou percentual acima de 65%. Para o critério de comparação da divergência plena (DPe), o percentual entre os *batches* 01 e 05 variou entre 2,60% e 5,90%, o que denota que, mesmo numa situação regular de balanceamento da razão regulatória (ora viés < variância, ora viés > variância), o percentual se iniciou em 3,80% no *batch* 01 e finalizou em 4,40% no *batch* 05, denotando uma divergência estimada menor que 6,00% nas cinco revisões.

E, finalmente, para o caso do critério de comparação com convergência parcial (CPa), os *batches* iniciais, 01 e 02, tiveram uma pequena distinção de percentuais para os três times, e uma nítida recuperação da razão regulatória para os *batches* 03, 04 e 05, com diferenças percentuais ínfimas em, no máximo, 1,00%.

De modo geral, nas duas línguas, o *batch* revalidado *CPe* teve um progresso percentual final de ~75,0%, representando que a revisão coletiva converge sob a sujeição comum da ética em cada membro; o *batch* revalidado DPe teve um regresso percentual final de ~2,5%, denotando que a diminuição da discordância tem relação com parâmetros pontuais de cada texto, necessitando uma análise e avaliação discursiva intensa, ao mesmo tempo que as relações político-público-

-sociais são inferidas para que, eticamente, eleja-se um critério de categorização a cada texto; O *batch* revalidado CPa manteve-se numa proporção entre ~5,0% – ~9,0%, sendo que a maioria de revisão era sobreposta à minoria, mas com identificação relacional sobre o viés direcionado nas reclassificações errôneas.

Alguns *inputs* apresentavam similaridades textuais, mesmo diferindo em sua integridade de itens lexicais, sendo necessário neste caso a implementação de *Augmentation*, para manter o discurso inicial da significação semântica, mas alterando itens pretextuais que convalidam o contexto do *input*, sendo uma referência recursiva somente nesta necessidade. Para *inputs* relacionados a categoria "*Not Urban*", pretendeu-se durante a revisão/reclassificação das bases uma identificação do nível de socialização pública em que o texto se encaixaria como característica urbana, denotando que tal texto deve ter um discurso que, mesmo tendo um lexical informativo sobre determinada temática, é dissociado de itens pré-textuais que o direcionem a alguma característica urbano-social.

Assim, mitigar *bias* significa que, mesmo tendo níveis elevados de revisão/reclassificação na convergência plena de categorias, seu viés final terá uma tendência minimamente associada a um senso comum, porém prevalecendo a normativas éticas sobre o entendimento sociomoral que cada ser-humano se responsabiliza no desenvolvimento de algoritmos que tratam com informações de interesses político-públicos mútuos (Chakravarthi, 2020; Hudley et al., 2020; Joshi et al., 2020; Liu et al., 2020).

Conclusões

Neste projeto de pesquisa, buscou-se revisar várias bases de dados de treino e testes com o propósito de mitigar e minorar os vieses pessoais em um modelo multimodal de classificação de categorias urbano-sociais. A caracterização contextual e dialógica de textos manifestados digitalmente com cada categoria tem relação direta com a subjetividade de cada indivíduo em seu respectivo texto, mostrando seus vieses construídos moral e cognoscentemente.

A conversão coletiva da moralidade e do caráter de convalidação linguística durante as revisões dos *batches* mostrou que a análise textual em PLN necessita de uma identidade racional às premissas e prerrogativas, *a priori*, do que se entende por um contexto político-público e sociomoral. Isto posto, durante o desenvolvimento de modelos de IA/ML que lidam diretamente com o público, o balanceamento entre viés e variância é necessário para seja que permissível e praticável uma atualização adaptável nos dados de treinamento e teste sob uma supervisão variável, diversa e heterogênea de reclassificação e revalidação.

Os referenciais teóricos da linguística discursiva, da construção da moralidade e das abordagens analíticas sobre viés/variância fomentaram os fundamentos que este projeto de pesquisa pudesse atingir assertivamente o objetivo da mitigação de *bias* que, mesmo sendo uma tarefa laboriosa, é de pauta algorítmica-social para manter a pluralidade e robustez em dados públicos.

Agradecimentos – Este projeto de pesquisa é financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp), pelo processo n.2019/19032-6.

Referências

ANGWIN, J. et al. *Machine bias*: There's software used across the country to predict future criminals and it's biased against blacks. ProPublica, 2017.

ASSIMAKOPOULOS, S. et al. Annotating for hate speech: The MaNeCo corpus and some input from critical discourse analysis. In: PROCEEDINGS OF THE 12TH LANGUAGE RESOURCES AND EVALUATION CONFERENCE, Marseille, p.5088-97, Marseille, France. 2020.

BAKHTIN, M. M. *Estética da criação verbal* (edição francesa Tzvetan Todorov). 6.ed. São Paulo: Editora MF, 2011.

_____. *Problemas da poética de Dostoiévski*. 5.ed. Rio de Janeiro: Forense Editora, 2018.

BASU, P. et al. *Multimodal Sentiment Analysis of #MeToo Tweets using Focal Loss* (Grand Challenge). 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM).

BLODGETT, S. L. et al. Language (technology) is power: A critical survey of "bias" in NLP. In: PROCEEDINGS OF THE 58TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, p.5454-76, Online. Association for Computational Linguistics. 2020.

BLODGETT, S. L. et al. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In: PROCEEDINGS OF THE JOINT CONFERENCE OF THE 59TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS and the 11th International Joint Conference on Natural Language Processing, Online. Association for Computational Linguistics. 2021.

BRAIT, B. (Org.). *Bakhtin. Dialogismo e construção do sentido*. Campinas: Editora da Unicamp, 2005.

_____. (Org.). *Bakhtin e o Círculo*. São Paulo: Contexto. 2009.

BRANDIST, C.; TIHANOV, G. *Materializing Bakhtin. The Bakhtin Circle and social theory*. London: MacMillan Press, 2000.

BRASIL/EBIA. *Estratégia Brasileira de Inteligência Artificial* (EBIA) em 07/2021 – Ministério da Ciência, Tecnologia e Inovações (MCTI) / Secretaria de Empreendedorismo e Inovação (SEI). 2021.

BROWN, A. et al. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In: PROCEEDINGS OF THE 2019 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, CHI '19, New York, p.1-12, New York, NY, USA. Association for Computing Machinery. 2019.

CAÑETE, J. et al. Spanish Pre-Trained BERT Model and Evaluation Data. In: Practical ML for Developing Countries Workshop (PML4DC) at Eighth International Conference on Learning Representations (ICLR). Addis Ababa, Ethiopia CFP2020, PML4DC at ICLR 2020.

CASTELLE, M. The linguistic ideologies of deep abusive language classification. In: PROCEEDINGS OF THE 2ND WORKSHOP ON ABUSIVE LANGUAGE Online (ALW2), Brussels, p.160-70, Brussels, Belgium. Association for Computational Linguistics. 2018.

CHAKRAVARTHI, B. R. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In: PROCEEDINGS OF THE THIRD WORKSHOP ON COMPUTATIONAL MODELING OF PEOPLE'S OPINIONS, PERSONALITY, and Emotion's in Social Media, Barcelona, p.41-53, Barcelona, Spain (Online). Association for Computational Linguistics. 2020.

DEVLIN, J. et al. *Bert: Pre-training of deep bidirectional transformers for language understanding*. S.l.: s.n., 2018.

FIELD, A. et al. A Survey of Race, Racism, and Anti-Racism in NLP. In: PROCEEDINGS OF THE 59TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS and the 11th International Joint Conference on Natural Language Processing, p.1905-25. August 1–6, 2021a.

FIELD, A.; PARK, C. Y.; TSVETKOV, Y. *Controlled analyses of social biases in Wikipedia bios*. Computing Research Repository, arXiv:2101.00078. Version 1. 2021b.

GILLANI, N.; LEVY, R. Simple dynamic word embeddings for mapping perceptions in the public sphere. In: PROCEEDINGS OF THE THIRD WORKSHOP ON NATURAL LANGUAGE PROCESSING AND COMPUTATIONAL SOCIAL SCIENCE, Minneapolis, p.94-9, Minneapolis, Minnesota. Association for Computational Linguistics. 2019.

HANNA, A. et al. Towards a critical race methodology in algorithmic fairness. In: PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, New York, p.501-12, New York, NY, USA. Association for Computing Machinery. 2020.

HILT, D. E.; SEEGRIST, D. W. *Ridge: a computer program for calculating ridge regression estimates*. Research Note NE-236. Upper Darby, PA: U.S. Department of Agriculture, Forest Service, Northeastern Forest Experiment Station. 7p. 1977.

HUDLEY, A. H. C.; MALLINSON, C.; BUCHOLTZ, M. Toward racial justice in linguistics: Interdisciplinary insights into theorizing race in the discipline and diversifying the profession. *Language*, v.96, n.4: p.e200–e235, 2020.

HUTCHINSON, B. et al. Social biases in NLP models as barriers for persons with disabilities. In: PROCEEDINGS OF THE 58TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, p.5491-501, Online. Association for Computational Linguistics. 2020.

JIANG, M. et al. *Transformer Based Memory Network for Sentiment Analysis of Web Comments*. IEEE Access: Special section on Innovation and Application of Intelligent Processing. DOI: 10.1109/ACCESS.2019.2957192, 2019.

JIANG, M.; FELLBAUM, C. Interdependencies of gender and race in contextualized word embeddings. In: PROCEEDINGS OF THE SECOND WORKSHOP ON GENDER BIAS IN NATURAL LANGUAGE PROCESSING, Barcelona, p.17-25, Barcelona, Spain (Online). Association for Computational Linguistics. 2020.

JOSHI, P. et al. The state and fate of linguistic diversity and inclusion in the NLP world.

In: PROCEEDINGS OF THE 58TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, p.6282-93, Online. Association for Computational Linguistics. 2020.

KIELAY, D. et al. *Supervised Multimodal Bitransformers for Classifying Images and Text*. arXiv:1909.02950v2 [cs.CL] 12 Nov 2020.

LAHIRE, B. *Formes Sociales Scripturales et Formes Sociales Orales. Une Analyse Sociologique de l''Echec Scolaire' à l'Ecole Primaire*. Lyon, 1990. Tese (Doutorado) – Université Lumière Lyon 2.

_____. *Culture écrite et inégalités scolaires*. Lyon: Presses Universitaires de Lyon. DOI : 10.4000/books.pul.12525. 1993a.

_____. *La raison des plus faibles*. Rapport au Travail, Ecritures Domestiques et Lectures en Milieux Populaires. Lille : Presses Universitaires de Lille. 1993b.

_____. Pratiques d'écriture et sens pratique. In: SINGLY, F. de; CHAUDRON, M. (Org.) *Identité, Lecture, Ecriture*. Paris: Bibliothèque Publique d'Information; Centre Georges Pompidou, 1993c. p.115-30.

LEINS, K.; LAU, L. H.; BALDWIN, T. Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In: PROCEEDINGS OF THE 58TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, p.2908-13, Online. Association for Computational Linguistics. 2020.

LEPORI, M. Unequal representations: Analyzing intersectional biases in word embeddings using representational similarity analysis. In: PROCEEDINGS OF THE 28TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, Barcelona, p.1720-8, Barcelona, Spain (Online). International Committee on Computational Linguistics. 2020.

LIU, H. et al. Does gender matter? towards fairness in dialogue systems. In: PROCEEDINGS OF THE 28TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, Barcelona, p.4403-16, Barcelona, Spain (Online). International Committee on Computational Linguistics. 2020.

MA, K. Artificial unintelligence: How computers misunderstand the world. *The Information Society*, v.35, n.5, p.314-15, 2018. DOI: 10.1080/01972243.2019.1655942.

MELLET, K. et al. A "democratization" of markets? Online consumer reviews in the restaurant industry. *Valuation Studies*, v.2, n.1, p.5-41, 2014. doi: 10.3384/vs.2001-5992.14215.

MOTHA, S. Is an antiracist and decolonizing applied linguistics possible? *Annual Review of Applied Linguistics*, v.40, p.128-33, 2020.

NOBLE, S. *Algorithms of oppression*: How search engines reinforce racism. New York: NYU Press, 2018.

O'NEIL, C. *Weapons of math destruction*: how big data increases inequality and threatens democracy. New York: USA C. Publishers, 2016. ISBN 9780553418835.

PÊCHEUX, M. *Semântica e discurso*: uma crítica à afirmação do óbvio. Ed. Unicamp, 1988.

_____. *O discurso*: estrutura ou acontecimento. Campinas: Pontes, 1990.

PIAGET, J. *Adaptation Vitale et Psychologie de l´Intelligence*: sélection organique et phénocopie. France: Hermann, 1974.

_____. *A epistemologia genética, sabedoria e ilusões da filosofia, problemas de epistemologia genética*. São Paulo: Abril Cultural, 1983.

_____. *Seis estudos de Psicologia*. 18.ed. Rio de Janeiro: Forense Editora, 1991.

_____. *O juízo moral na criança*. São Paulo: Summus, 1994.

SOUZA, F., NOGUEIRA, R., LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9TH BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS, BRACIS, Rio Grande do Sul, Brazil, October 20-23, 2020.

TESTUGGINE, D. et al. Supervised Multimodal Bitransformers for Classifying Images and Text. *arXiv:1909.02950v2* [cs.CL] 12 Nov 2020.

VASWANI, A. et al. Attention is all you need. In: 31st CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NIPS 2017), CA, USA. *arXiv preprint*: *arXiv:1706.03762v5*. 2017.

YOON K.; DENTON, C.; HOANG, L. Structured attention networks. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, s. l., 2017.

RESUMO – O referido projeto se caracteriza nas implicações relacionais do desenvolvimento sociomoral da teoria psicogenética em Piaget sobre a construção cognoscente da ética nos vieses pessoais e em referenciais da dialética discursiva na linguística. Foram parametrizados a dados funcionais de treinamento e teste em um classificador de categorias urbano-sociais em uma abordagem analítica textual por Processamento de Linguagem Natural (PLN), e baseado no mecanismo de atenção adaptada Transformers. Nessa perspectiva, desenvolveu-se uma metodologia de mitigação de viés para a reestruturação do crivo e critério que *datasets* multimodais são retreinados, retestados e reavaliados. Finalmente, verificou-se e validou-se a heterogeneidade da ética comum coletiva humana, sobre inferências interpretativas, *insights* e tendências sociais reais que a relação cidade/cidadão aborda o "*social sensing*" na identificação de problemas público-sociais.

PALAVRAS-CHAVE: Mitigação de viés, *Social sensing*, *Transformers*, Análise de textos em PLN, Classificação de textos.

ABSTRACT – This research project is based on the relational implications of the sociomoral development of Piaget's psychogenetic theory on the cognition construction of ethics in personal biases as in references of discursive dialectics in linguistics. Functional data from training and testing were parameterized in an urban-social category classifier in a textual analytical approach by Natural Language Processing (NLP) and based on the Transformers adapted attention mechanism. In this perspective, a bias mitigation methodology was developed to restructure the convergence criteria in which multimodal datasets were retrained, retested, and reevaluated. Finally, the heterogeneity of the common collective human ethics was verified and validated, over interpretive inferences, insights, and real social trends, whereby the city/citizen relation addresses the "social sensing" in the identification of public-social problems.

KEYWORDS: Bias mitigation, Social sensing, Transformers, NLP text analysis, Text classification.

Luciano C. Lugli é bacharel em Engenharia da Computação (2008), mestre em Engenharia Mecânica/Mecatrônica (2011) e doutor em Engenharia Mecânica/Mecatrônica (2016) pela Escola de Engenharia de São Carlos da Universidade de São Paulo. Engenheiro de Dados Sênior (desde 2021) na Daoura Research – São Paulo, SP, Brasil. @ – luciano.lugli@daoura.ai / https://orcid.org/0000-0002-9065-9639.

Daniel Abujabra Merege é bacharel em Sistemas de Informação pela Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (2010), mestre em Engenharia da Computação pelo Instituto de Pesquisas Tecnológicas do Estado de São Paulo (2016). Co-fundador e CEO (desde 2016) na Daoura Research – São Paulo, SP, Brasil. @ – daniel@daoura.ai / – https://orcid.org/0000-0002-9232-9270.

Rafael Pillon Almeida é bacharel em Ciência da Computação pelo Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (2012). Head de Tecnologia (desde 2018) na Daoura Research – São Paulo, SP, Brasil. @ – rafael@daoura.ai / https://orcid.org/0000-0003-0558-276X.

Recebido em 26.1.2023 e aceito em 27.2.2024.

[I] Universidade de São Paulo, Escola de Engenharia de São Carlos, Daoura Research, São Paulo, Brasil.

[II] Instituto de Pesquisas Tecnológicas do Estado de São Paulo, Daoura Research, São Paulo, Brasil.

[III] Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, Daoura Research, São Paulo, Brasil.