DOI: 10.1590/s0103-4014.202438111.019

Bias mitigation of multimodal datasets in an urban-social category classifier

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

Introduction

Piaget (1974; 1983), a reference in pedagogical psychogenesis, dedicated his research to the knowledge construction, which he called genetic epistemology, stating "[...] the development of several variables of knowledge, since its elementary mode, and following its epistemic evolution from birth to the subsequent levels, uniquely builds the ethics in human beings" (Piaget, 1983, p.3). And also, according to Piaget (1991, p.23) the moral development is extensive and asserts that "[...] morality is endowed with a system of rules, and the ethical essence of morality must be sought in the respect that the individual acquires for these rules."

In this Piagetian perspective, this project aims to develop a bias mitigation methodology on multimodal datasets (Yoon, 2017; Jiang et al., 2019; Testuggine et al., 2020) intended to an urban-social category classifier, in which the social listening of assorted text manifestations about municipality issues, demands an implicative interpretation of the (inter)locutor on such a specific subject. The identification of subjectivity on persons' speech is based on the linguistic parameterization from the theory of literacy (Bakhtin, 2011; 2018; Pêcheux, 1988; 1990), and it is required as reliable for the contextual convergence of urban-social categories in the inferred texts.

Although, stratifying training data and how to sense testing data must be an heterogeneous task regarding the individual subjectivity of textual manifestations, as the same time an homogeneous task regarding the subjection of collective convergence to the social paradigms that distinguish the character in the cognitive-human interpretation of these textual manifestations. Thus, to collectively converge the moral and social values in Machine Learning (*i.e.*: and in this case in NLP – Natural Language Processing – textual analysis) it is necessary a rational identity to the political-public and socio-moral precepts wherein software developers can cross on a minimal (yet reasonable) bias and permissible the plurality of data whether in training or testing (Mellet et al., 2014; O'Neil, 2016; Noble, 2018; Ma, 2018).

Mitigating bias is a task that, even initially laborious, is necessary when it is intended to infer the details on urban-social categories whose significance is minimally present in their semantics (*e.g.*: on verifying if a text is *political*, but has some features from *governance*, and also financial factors from *economy*).

Objectives

This project aims to relate the psychogenetic premises of sociomoral development and discursive linguistics in bias mitigation of training and testing data, regarding an urban-social category classifier, based on Transformers adapted attention mechanism (Vaswani *et* al., 2017) – a ML architecture that infers directly and indirectly on the parallel data processing distributed in tokens, then contextualized by minimal relational convergence. It is intended to verify the existing ethics integrity criterion in which multimodal datasets support and sustain the erroneous classification of categories (false negatives/false positives), thus impairing the parameters of final precision classification.

Theoretical references

To elucidate text manifestations describing public common problems related to a municipality, it demands an inherent characterization of the interlocutor person on such topic, in an attempt to identify the subjectivity in the speech (*who speaks to, what he/she speaks to , from where he/she speaks to, to whom he/she speaks, ...*) (Bakhtin, 2011). Thereby, a linguistic/literacy academic support is necessary to normalize its respective matter, becoming both *technical-technological* and *linguistic-social* contextual composed.

Thus, in this section, the theoretical references that ratify the moral linguistic base in which bias is built from different cultural contexts were described. When dealing with an NLP analysis focused on public interaction texts from people on social nets virtual platforms, the context shall consider an interpretation that validates the statements structured schematically by the corpora, vocabulary and semantics (Brait, 2005), enabling the ML neural network to identify all lexicon details in one category to another.

Mitigating bias in this perspective, confirms the convergence in which both the contextual meaning of each person responsible for feeding the *texts-categories* relation in datasets, as also the criterion to differentiate the urban-social classes, is inherently related to comprehensive development (*e.g.*: morality, cognition, consciousness, behavior, …) (Angwin et al., 2017; Assimakopoulos et al., 2020; Blodgett et al., 2020).

From Piaget's premises (1994, p.23), the moral development is vast, it does not refer to just obeying certain rules, but to understanding the reason why the subject obeys them, therefore, he ensures that "[...] all morality consists of a system of rules, and the essence of all morality must be sought in the respect that the individual acquires for these rules."

His research (Piaget, 1994) validated the process of moral development is a tricky task, commonly in the social context, in which the prerogatives of

development and cognitive social interaction overlap with morality. Even based on documental references wherein instigate pedagogical methodology for moral development since early childhood education, the interpersonal relationships as the integral development of the human being would be discreditable by antisocial behaviors emerged by his/her bias.

Thus, a discursive analysis is necessary to check not only the subjacent speech of the speaker but also his/her morality between who speaks to and what it is spoken to. So, in these assumptions, such analysis is directly related to a list (NN classes) of urban-social categories regarding the city-to-citizen issues, *i.e.*: the management of services employed in a municipality.

To inspect these urban-social categories involved in each contextual text semantics, it is necessary to ground the baseline approach with the linguistic subjectivity (Lahire, 1990; 1993a) within intrinsic parameters of *Transformers adapted attention mechanism* architecture, which discretizes the significance and semantics of the whole word lexicon (*n*-grams) through a customized model named *parallel relational integration of linguistic construction*.

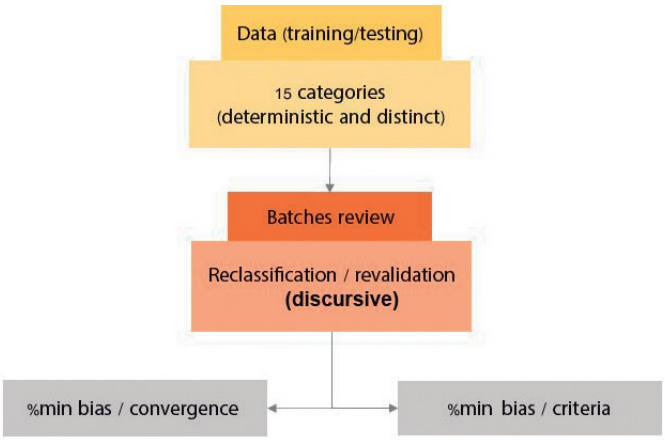
Understanding the underlying language in social networks by Bakhtin

Language only exists as a function between *who* speaks to, *what* is spoken, and *to whom* it is spoken the contextual communication of digital manifestations. Teaching, learning, and using the language necessarily pass through the subject, the social relations entity, and the person responsible for the composition fundamentals of the related discourse. This subject uses the previous information to formulate his/her speeches and statements. Besides, an underlying utterance is modulated by the speaker for the social and cultural context, otherwise it will not be understood (a reference to the public mischaracterization for the "*Not Urban*" category) (Bakhtin, 1929/2011; 2018; Brait , 2009; Brandist & Tihanov, 2000).

In this dialogical relationship between interlocutors in the social environment, in which the verbal and non-verbal influence settle the process of utterances' construction, the interaction through language befalls in a context which everyone is exemplary equal (homogeneity to manifest). Those who eligibly enunciates, withal address the proper understandable undergoing messages (*i.e.*: commenting directly on an urban-social category). Yet, those who latently listen to, interpret, and parameterize the meaning of the utterance (through its systemic inference – *e.g.*: NLP).

Method

The methodology for bias mitigation project purpose is structured in training and testing databases to a referential ML multimodal model intended to the classification of 15 different categories. A sequential diagram of such proposed method is described below.



Source: Elaborated by the author.

Figure 1 – Sequential diagram of bias mitigation strategy on the training and testing databases.

To contextualize the related problem related to bias in NLP, an urban-social category classifier was implemented, based on Transformers adapted attention mechanism (Vaswani et al., 2017), which is parameterized according to the adapted tunning of *MultiLabeling* task classification over the *simpletransformers* library, in Portuguese language (*HuggingFace / BERTimbau* pre-training + tokenizer based on *neuralmind/bert-base-portuguese-cased*) (Souza et al., 2020), and in Spanish language (*HuggingFace/BETO* pre-training + tokenizer based on *dccuchile/bert-base-spanish-wwm-cased*) (Cañete et al. 2020), both *HuggingFace/*BERT-based pre-trained models (Devlin et al., 2018).

The urban-social category classifier is used as a multi-class classification reference, as it reflects the public-social context in a municipality, in which to be correctly categorized, the text shall present suitability and significance with an urban-social functional factor, and respond to: Is this text related to any aspect of public management? What repercussions does the contextual content of this text relate on management and public services? Does this text regard improvements on public and social services? Does this text provide a socially participatory management in the city (and the citizens)?

In this multimodal NN model, the classifier relies on 15 urban-social categories, defined and described to a social-urban factor/function (*"Culture and Recreation", "Economy", "Education", "Environment", "Governance", "Healthcare", "Housing/Social Assistance", "Mobility", "Politics", "Security", "Tourism", "Urban Infrastructure", "Waste Management", "Water and Sanitation"*, and, when a text mischaracterizes any of the previous categories, without the urban-social intelligibility, as also the misclassification arguments for categories, it is defined as "*Not Urban*"). A text is classified according to its corpus (content and context), and subjective semantics (expressions/excerpts that directly and

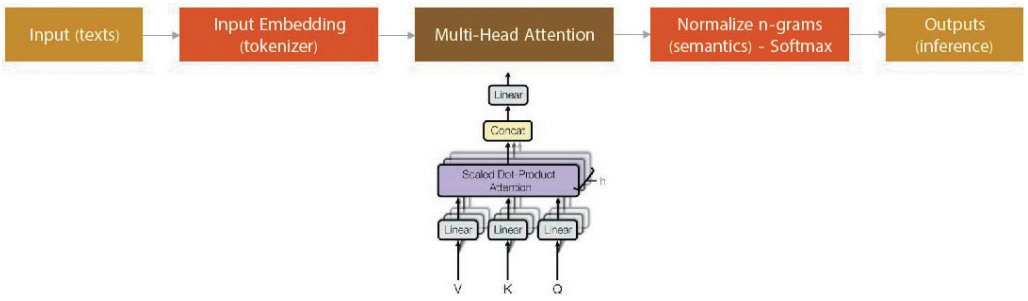
indirectly relate the corpus) in a class (with maximum %confidence for a category) (*e.g.*: a text is classified only as "Education"), or in more classes (with a minimum %confidence for more than one category) (*e.g.*: a text is partially classified as "Education" and "Politics", providing linguistic/lexical characteristics of these two categories).

Transformers adapted attention mechanism for multi-class text classification

Transformers is a neural network architecture that prioritizes "attention" parallelism, rather than recurrence, which in the *seq2seq* context, the intermediate attention layers greatly improve the performance of neural networks (Yoon; 2017; Basu, 2020) in parallelize the lexicon in NLP.

Such parallelization provides a textual analysis in a correlational way, which means that the *n*-grams (*e.g.*: words within, subject, verbs, object) are analyzed in the lexical structural scheme in: "*one by one*" and "*one to one*", inferring wherefore into functions that feature whenever a text is being classified as previously intended (Jian et al., 2019; Kielay et al., 2020).

A priori, this classifier resulted in erroneously biased criteria predictions with correlational connotation, restraining the prediction percentages of the base training categories (*e.g.*: category 1 is FN and category 2 is TP). This is why the bias mitigation problem had to be performed.



Source: Elaborated by the author.

Figure 2 – General diagram of Transformers adapted Attention mechanism.

Bias mitigation – review, revalidation, and reclassification

Throughout the 15 categories, the batches were started with a balanced increasing amount of texts/category, so the solution was to quantify the training texts and qualify the contextual content (*i.e.*: textual corpora), thus mitigating the bias between them.

The training sets were initially divided into partial-periodically reviews, distributed in batches with a specific quantity during the week. Each subsequent weekly partial set had a text/category distribution that depended on the current review, thus providing continuous ratio balancing then providing a homogeneous quantity. As for quality, there were texts with contextual utterance understanding related to the semantics of linguistic subjectivity (Lahire, 1990; Gillani;

Levi, 2019; Hanna et al., 2020; Leins et al., 2020) along with the intrinsic parameters of the adapted attention mechanism, which discretizes the significance and meanings of *n*-grams. This means that texts should have paragraphs, phrases or sentences that shall rely on lexicon of tags related to their respective category.

Three review teams were set for reclassifying the texts in each batch: '*T1*' was a diversified team with AI/ML and NLP specialists; '*T2*' was a diversified team with specialists in linguistics (graduated in Literature/Linguistics); '*T3*' was a heterogeneous team representing social class/community minorities.

The reviews were structured among *absolute convergence* (total conformity review) – *AC*, *partial convergence* (major conformity review) – *PC*, and *absolute divergence* (total unconformity review) – *AD*. In this last type, a detailed review was carried out text by text to identify the various biases, then discussed under the linguistic discourse of Lahire analytical approach (Lahire, 1993b; 1993c), as also under the epistemic psychogenesis on ethics, which the arguments were endorsed to their respective classifications, thus assenting to a final category criterion by the members.

In this batch review procedure, it was intended to differentiate the varieties of biases that developers shall prevent: *algorithm bias, sample selection bias, prejudice bias and measurement bias* (Brown et al., 2019; Field et al. , 2021a; 2021b).

Algorithm bias: refers to the ML algorithm's own property on the model trend performance between training and testing. Therefore, a balance must be settled between this bias and the data variance (balancing the quantity and quality of texts);

Sample selection bias: refers to the representative quality of sample accuracy on which the data is trained. Samples from heterogeneous populations in terms of quality must be used to validate their homogeneous representativeness over the sample size;

Prejudice/preconception bias: refers to the ethical content about "consciousness, behavior and cognition" relationship that developers have in their political-social conceptions. Mitigating this bias requires restrictions on the sociomoral matters that data were classified;

Measurement bias: refers to the measurement/metric that the data is performed, with systematically skewed results. There must be a balance between data acquisition and analysis to standardize different natures of data.

Based on this methodology, it is relevant to gather the systemic legislation information, regarding the Brazilian AI Strategy (Brazil/EBIA, 2021), in which it describes in its chapter "Legislation, Regulation and Ethical Use", what is the implication to identify biases involved in decision-making interpretations (pattern recognition) when appropriate to the ethical matter in ML, based on discussions about reliable and AI-centered systems in the *human-machine* relationship (*trustworthy H-M AI*).

Results

The results of review/reclassification of training data show an assertive accuracy of ~90.0%, post bias mitigation, considering cross-contextual convergence, given data discrimination by parallel criterion inference. This identifies a progression of batch revalidation percentages.

Moreover, the validation of modeled data was based on the ratio between bias and variance. A low bias means that the model learns, but in an adjusted variance, which it is impossible to discriminate data other than those associated, *a priori*, in the training dataset (Blodgett et al., 2021; Castelle, 2018; Hutchinson, 2020; Jiang et al., 2020; Lepori, 2020; Motha, 2020).

To maintain the regulatory ratio in the model, the variance is reduced by the heterogeneity of relational data of each bias (each review team), managing to converge and generalize the testing dataset, *a posteriori*, by a ridge regression (Hilt et al., 1977) parametrization in the reviewed data.

For Portuguese language model there were 7 training/test batches and for Spanish language model there were 5, since each database refers to a distinct set of quantity per class, *tokenizers* on the intended datasets, and targeted strategies for each language.

Tables 1 and 2, present the review convergence and criterion percentages defined by the three sorted teams 'T1', 'T2' and 'T3'. When these three teams match, there is *absolute convergence (AC)*; when one of the teams is not equal to the others, there is *partial convergence (PC)*, in which three situations of different biases were reviewed; and when none of the teams match to each other, there is *absolute divergence (AD)*. Line 01 shows the *AC* (teams T1, T2, T3 have the same biases); Line 02 shows the *AD* (teams T1, T2, T3 have different biases); Lines 03–05 show partial convergence – *PC* (three situations in which two teams have the same bias but differ to each other).

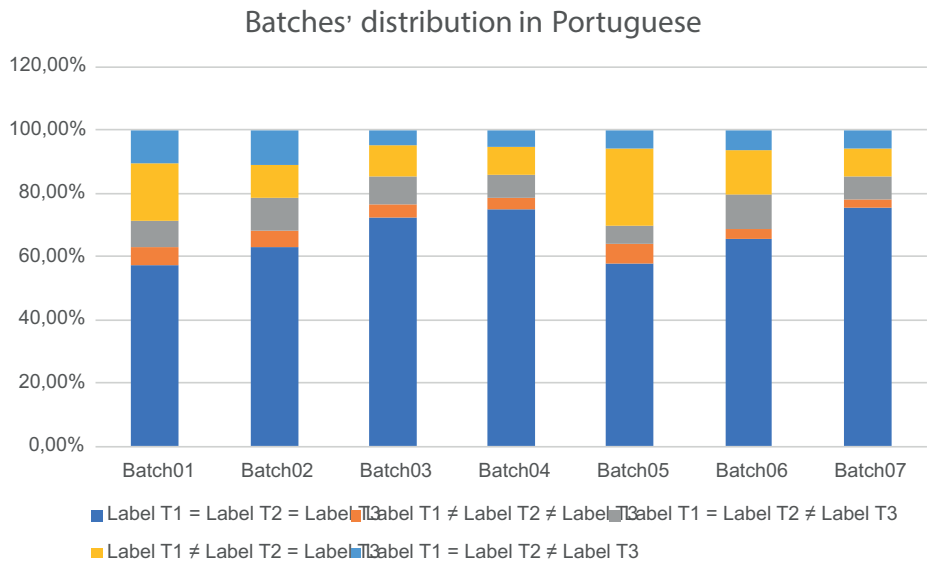
Table 1 shows the percentage results on batches 01–07 and the 5 correspondences of convergence/criteria for Portuguese language.

In Portuguese, the data in Table 1/Figure 3 show progress in the percentage for *absolute convergence (AC)* on batches 01 and 04 (from 57.20% to 74.80%). Batch 05 presented different data from the previously trained bias, promoting heterogeneity in data reclassification (*variance* > *bias*), and the comparison criterion among teams T1, T2 and T3 was reestablished, and evolved normally as more data were reviewed between batches 05 and 07. At the end of the 7 batches review, the *AC* obtained a percentage of 75.40%. For the absolute divergence comparison criterion (*AD*), the percentage between batches 01 and 07 decreased significantly, and ended at 2.60% minimum. Even in the same situation of the regulatory ratio (*bias* <> *variance*) from batch 05, this percentage is below the estimated review, not reaching the maximum 7%, and denotes that the biases of the three teams were concise and convergent in the categorization.

Table 1 – Distribution of batches and respective percentages for each discrimination defined by the review strategy (Portuguese language)

Comparison criteria	*Batch*01	*Batch*02	*Batch*03	*Batch*04	*Batch*05	*Batch*06	*Batch*07
Label T1=Label T2=LabelT3	57,20%	63,20%	72,20%	74,80%	57,60%	65,40%	75,40%
Label T1≠Label T2≠Label T3	5,60%	4,80%	4,40%	3,60%	6,20%	3,40%	2,60%
Label T1=Label T2≠Label T3	8,60%	10,40%	8,80%	7,60%	6,20%	11,00%	7,40%
Label T1≠Label T2=Label T3	18,00%	10,80%	9,60%	8,80%	24,40%	14,00%	9,00%
Label T1=LabelT2≠Label T3	10,60%	10,80%	5,00%	5,20%	5,60%	6,20%	5,60%

Source: Elaborated by the authors.



Source: Elaborated by the authors.

Figure 3 – Distribution of batches and respective percentages in Portuguese.

Finally, for the case of the comparison criterion with partial convergence (PC), two batches had predictable results due to the unbalanced regulatory ratio (bias > variance), batch 01, because it was the initial review, and batch 05, in balancing the regulatory ratio (bias < variance). For the other batches review, the percentages of partial convergence among the three teams remained in a plausible and permissible proportion.

Table 2 shows the percentage results on batches 01–05 and the 5 correspondences of convergence/criteria for Spanish language.

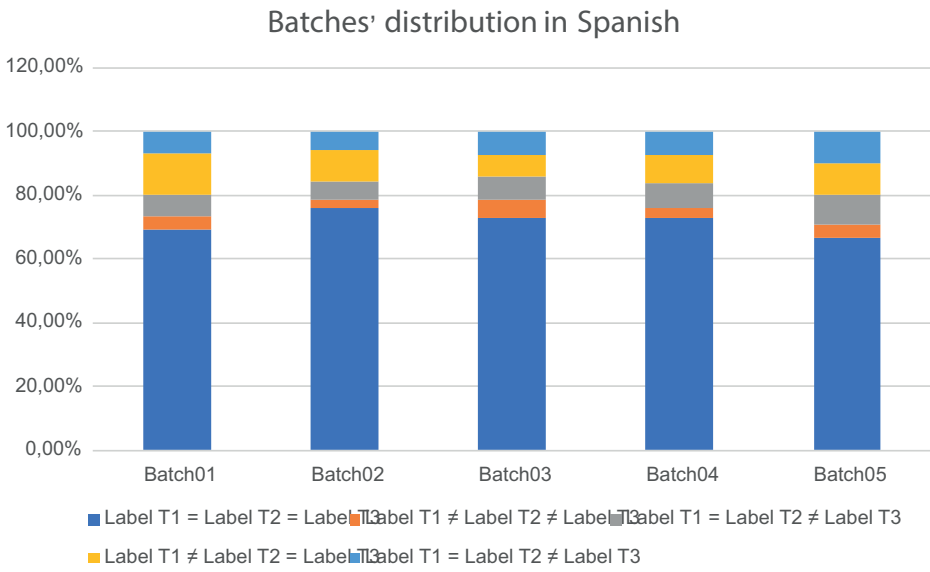
In Spanish, the data in Table 2/Figure 4 show different progress than in Portuguese. Initially, only 5 batches were fixed, and the first (batch 01) was significantly smaller than the other batches, as it was a validation test with more

critical categories (e.g.: many classes erroneously in 'Not Urban'). Consequently, the other reviews were promoted in the characterization of batch 01, and in the absolute convergence comparison criterion (AC), the percentage varied between 66.40% and 75.80%, and the last reviewed batch presented a percentage above 65%. For the absolute divergence comparison criterion (AD), the percentage between batches 01 and 05 varied between 2.60% and 5.90%, which denotes that, even in a regular situation of balancing the regulatory ratio (sometimes bias < variance, sometimes bias > variance), the percentage started at 3.80% in batch 01 and ended at 4.40% in batch05, denoting an estimated divergence of less than 6.00% in the five revisions.

Table 2 – Distribution of batches and respective percentages for each discrimination defined by the review strategy (Spanish language).

Comparison criteria	Batch*01*	Batch*02*	Batch*03*	Batch*04*	Batch*05*
Label T1=LabelT2=Label T3	69,50%	75,80%	72,70%	73,00%	66,40%
Label T1≠LabelT2≠Label T3	3,80%	2,60%	5,90%	3,00%	4,40%
classe T1=Label T2≠Label T3	6,70%	6,00%	7,10%	7,80%	9,60%
Label T1≠LabelT2=Label T3	13,30%	10,00%	7,10%	8,80%	9,80%
Label T1=Label T2≠Label T3	6,70%	5,60%	7,30%	7,40%	9,80%

Source: Elaborated by the authors.



Source: Elaborated by the authors.

Figure 4 – Distribution of batches and respective percentages in Spanish.

And, finally, for the case of the comparison criterion with partial convergence (PC), the initial batches, 01 and 02, had a small difference in percentages for the three teams, and a clear recovery of the regulatory ratio for batches 03, 04 and 05, with minor percentage differences of a maximum of 1.00%.

In both languages, the *AC* revalidated batch had a final percentage progress of ~75.0%, representing that the collective review converges under the common subjection of ethics in each member; the *AD* revalidated batch had a final percentage return of ~2.5%, denoting that the decrease in disagreement is related to specific parameters of each text, requiring an intense discursive analysis and evaluation, at the same time as political-public-social relations they were inferred so that, ethically, a categorization criterion can be chosen for each text; The *PC* revalidated batch remained in a proportion between ~5.0% – ~9.0%, with the majority review overlapping the minority review, although with relational identification of the bias directed at erroneous reclassifications.

Some inputs presented textual similarities, even differing in their integrity from lexical items, being necessary in this case the implementation of *Augmentation*, to maintain the initial speech of the semantic meaning, but changing pre-textual items that validate the context of the input, being a recursive reference only in this need. For inputs related to the *"Not Urban"* category, it was intended during the review/reclassification to identify the level of public socialization in which the text would fit as an urban characteristic, denoting that such text must be discursive and dissociated from pre-textual items that relate it to some urban-social characteristic.

Thus, mitigating bias means that, even with high levels of review/reclassification in the absolute convergence of categories, its final bias will minimally be associated with common sense, but prevailing ethical norms over the socio-moral understanding that each human being is responsible for in the development of algorithms that deal with information of mutual political-public interests (Chakravarthi, 2020; Hudley et al., 2020; Joshi et al., 2020; Liu et al., 2020).

Conclusions

In this research project, a systematic review of training and testing datasets was performed to mitigate and reduce personal biases in a multimodal model for classifying urban-social categories. The contextual and dialogical characterization of digitally manifested text is directly related to the subjectivity of each individual in their respective category.

The mutual conversion of morality and the character of linguistic lexicon validation during batch reviews revealed that textual analysis in NLP specifies a rational identity to the premises and prerogatives, *a priori*, of what is underlying understood by a political-public and sociomoral context. Thereby, during the development of AI/ML models, the balance between bias and variance is necessary to enable an adaptive addition to training and testing data, based on heterogeneous supervision for reclassification and revalidation.

The theoretical references of discursive linguistics, the construction of morality and analytical approaches on bias/variance fostered the foundations that this research project assertively achieved on mitigating bias, which, even though it is a laborious task, is also mandatory as an algorithmic-social agenda to maintain plurality and robustness in public data.

Acknowledgments – This research project is funded by the Research Foundation of São Paulo State (Fapesp), under process n.2019/19032-6.

References

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 qualitati-

ve 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.

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.

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 Lin-

guagem 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.

Luciano C. Lugli has a B.S. degree in Computer Engineering (2008 – Northern São Paulo University), amaster's degreein Mechanical Engineering (2011– SãoCarlos School of Engineering / University of São Paulo) and a PhD in Mechatronics Engineering (2016 – São Carlos School of Engineering / University of São Paulo). He is a Senior Data Scientist / Engineer at Daoura Research since 2021 – São Paulo, SP, Brazil.
@ – luciano.lugli@daoura.ai / https://orcid.org/0000-0002-9065-9639.

Daniel Abujabra Merege has a B.S. degree in Information Systems (2010 – School of Arts, Sciences and Humanities / University of São Paulo), a master's degree in Computer Engineering (2016 – Institute of Technological Research of the State of São Paulo). He is the Co-founder and CEO at Daoura Research since 2016 – São Paulo, SP, Brazil.
@ – daniel@daoura.ai / https://orcid.org/0000-0002-9232-9270.

Rafael Pillon Almeida has a B.S. degree in Computer Science (2012 – Institute of Mathematics and Computer Science/ University of São Paulo). CTO at Daoura Research since 2016 – São Paulo, SP, Brazil. @ – rafael@daoura.ai /
https://orcid.org/0000-0003-0558-276X.

Received on 26.1.2023 and accepted on 27.2.2024.

[I] University of São Paulo, São Carlos School of Engineering, Daoura Research, São Paulo, Brazil.

[II] Institute of Technological Research of the State of São Paulo, Daoura Research, São-Paulo, Brazil.

[III] University of São Paulo, Institute of Mathematics and Computer Science, Daoura Research, São Paulo, Brazil.