Elimination of Bias in Sentiment Analysis

Nayan Arora
Faculty of Science and Technology
University of Canberra
Canberra, Australia
u3249907@uni.canberra.edu.au

***Abstract*—The existing Sentiment Analyses algorithms have a deep bias embedded into it. It fails to provide the most accurate information to users as it does not instill into its computational analyses the consideration for an individual's race, religion, age, gender, ethnicity, local languages spoken and the geographical location. With access to internet made easier by the day there is an ever-increasing demand for an accurate language processing model that can detect all biases for all user activity online. Thus, increasing the accuracy of computational human-emotion recognition using convolutional neural network methodologies. This research helps build an extensive dataset by organizing all information collected through focus groups, questionnaires, case studies and one on one interviews into different categories of bias. Different steps are added in the data collection processes to validate and evaluate the data to ensure and maintain quality standards. With consent, an agreement is signed by all participants in this research that helps maintain ethical standards required to use all information provided and collected. On successful implementation and execution of this research we expect to provide the stakeholders and government entities with increased user activity on their platforms by building trust to find less hate speech and biased information on their platforms and provide a pathway to future research opportunities.**

***Keywords*—convolutional neural networks, AI, NLP, biased dataset, user experience**

I. Introduction

Sentiment Analyses has been widely used and implemented across various ICT industries in the past ten years. Most innovations majorly took place in English speaking countries which is also why English has become the globally accepted language for all formal communication. With all technological advances and increasing affordability, technology has spread to all types of communities across the globe. Even the poorest of the poor has access to technology through their smartphones as they become cheaper by the day. This has increased the demand for more accurate locally built sentiment analyses systems that could be implemented across the web to help not just the ICT industry but the government as well. It is apparent that there exists a racial bias in hate speech detection and analyses [2]. It is also found that bias exists based on gender in the existing systems [3]. Even age-related bias has previously been proven to spread misinformation [6]. Sentiment Analyses algorithms have helped generate extensive revenue in various fields but has also been proven to be biased [7]. Hence, with the increasing use of technology in the age of artificial intelligence, we need to have more accurate sentiment analyses algorithms to understand human emotions and the feelings behind those emotions with the best possible accuracy.

II. Methodology

A. Planned Research Design

The initial information and scholarly articles were found using the keywords 'neural network', 'convolutional networks', 'sentiment analyses', 'language barriers' and 'racial bias'. A complete quantitative study has been previously conducted to achieve accurate results for sentiment analyses for twitter [7]. Although an accurate model was developed, the results and methodologies used to build the model fail to add considerations for language, race, religion, gender, age, and other possible barriers in understanding human emotion, that adds a lot of bias to it.

Hence, the new overall study would follow a mixed approach where both qualitative and quantitative data collected would provide results to understand and analyze, if further quantitative research is feasible for developing a new algorithm. So, this initial research paper is primarily focused on a core qualitative understanding and distinction of bias in sentiment analyses.

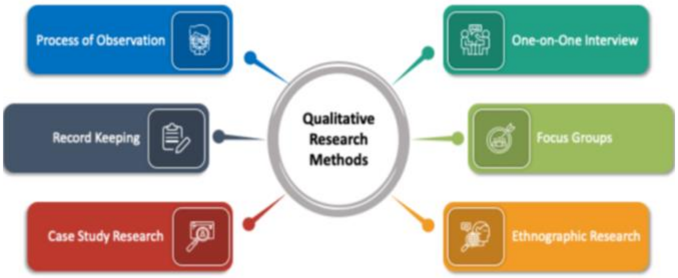


Fig. 1. Qualitative focused study design

By conducting a deep qualitative analysis using various methods, this research aims to identify the existing gap and answer the following research questions cohesively:

i. How can the current algorithm be altered to prevent from learning biased data and predicting the wrong human emotion?

ii. The advantages of having a convolutional neural model training technique that will consider all languages spoken around the world.

iii. How can current sentiment analysis or recommendation techniques be improved to include and understand a user's race, age, gender, or religious background to potentially eliminate all related bias?

iv. How are the current algorithms biased in tracking user activity and sentiments online?

Developing an insight on the feasibility of developing a new algorithm that can help eliminate all bias is the overall objective of this qualitative analyses. Further quantitative research studies would follow this study to develop the actual algorithm based on the results this research generates.

A clear focus will help us to generate and develop a precise dataset that infers and provides a deep understanding of all types of bias present in different communities globally. The steps needed to generate such a dataset is to produce an extensive literature review about the types of bias, followed by how they affect the general population in that country - positive, negative, or neutral. Focus groups from parts of the world provide useful data on the most spoken and understood language in their countries. A case study will be conducted to understand, analyze, and verify if the data that presents said bias is accurate. This will help prove the quality of the data we collect which will in turn help promote the following quantitative algorithm improvements. The methods to conduct this case study is discussed in detail in this research. One on one interviews will be conducted past this phase with people affected from such cultural biases in the past. All this information compiled will provide the basis for further research.

B. Planned Sample

Using our understanding from [1] and [3] we know that we would need to collect a lot of data both quantitative and qualitative from all resources to examine our hypothesis for any given bias. Since sentiment analyses is a language processing model it requires enormous amounts of data. To be able to review and analyze such huge amounts of data, we need a deep understanding of existing technologies that can help us in this venture. Regardless, there are techniques to discussed in [7] that we will implement and collect new data and use that to find matches in existing data. This will help distinguish our data into bias categories as age, gender, racial, religious, and cultural.

Furthermore, the numerous numbers of people participating in focus groups and one on one interviews will act as samples and help build our data. To maintain quality our team conducting these meetings will review the data at the end of each event to ensure data going into the overall categorial set is valid and true, this is further discussed in this research.

People will be chosen from around the world who have previously experienced any trauma because of racial, gender or any abuse. A focus would be kept on all individuals who speak two or more languages to be able to capture their understanding of different cultures. Moreover, a deep detailed case study will entail sampling information from students and teachers at universities worldwide. Only those universities that are globally recognized for their diverse student population will be allowed to participate based on the review of their expression of interest.

More data will be produced through randomly selected individuals from various social classes. The rich, the poor and the middle class all might have a different perspective of hate speech and it is crucial to capture that in our data.

Further a vast amount of data will be collected from social media platforms. Some of these bigger firms like twitter and Instagram will be requested to provide flagged data from the past five years, which will further improve our training data.

C. Planned Instrument for data collection

Since most new data collected is via human participation. This research will use a variety of instruments to collect our data.

Focus groups will be used to generate accurate categories for types of bias in sentiment analyses. Different focus groups will be conducted in different countries. Most discussions in these focus groups will be held around what is considered a bias and the kinds of bias that can be harmful for the general society. All focus groups will involve at least ten participants who hail from different cultural backgrounds, speak different languages, and have different social statuses in the society but are all living in the same city/geographical region. The team conducting focus reviews, overviews all information collected to find repetitions in results.

A case study will be conducted which will help capture if this research needs to produce more quality related restrictions. The case study developed will help to capture if information collected through focus groups is accurate. By seeking participation from people with proven misconducts in the past, the case study will try to ensure and approve of focus group results and further add all information collected into different types of bias. The case study will seek participation from people who may be in prison, hence would require government approval and consent.

Based on the results from this case study, further set of questions will be developed as a Questionnaire. This will be produced with a deep understanding of the current position of our research. All questions will be produced to capture the idea of positive, negative, or neutral bias in various statements or sayings which are developed using information from focus groups and the case study. Once the set of questions is finalized, it will be peer reviewed extensively. Then it will be handed out to people from different ethnic and cultural backgrounds of all ages worldwide.

Furthermore, university and student participation will be based on an expression of interest basis. As this research entails a possibility of a further quantitative research - which could result in the development of a new algorithm - which can be patented, we would require all university and student participants to sign a non-disclosure agreement.

Based on the dataset built by now, one on one interviews will be conducted with renowned individuals with proven academic or social history. For example, Greta who is a climate activist, prime ministers of different countries, Bill Gates who has been running multiple non-profits for several years and similar individuals will be invited to these interviews with a sole aim to capture their understanding of our work. Based on their reviews additions or subtractions will be made to our overall progress in this research.

D. *Planned Variables in the study*

To start with, we have found through our literature review the existing variables that have been studied before [4]. These variables will also be used in developing the instruments of our data collection.

i. Age (quantitative)

ii. Gender (qualitative)

iii. Race (qualitative)

iv. Religion (qualitative)

v. Ethnicity (qualitative)

But most of these variables will be implemented in the process of algorithm development in the next phase of this research and only help this current research to form distinctions categories for all other data we collect. There are more variables which are more specifically used in this research paper, produced by the instruments used for our data collection. They are summarized as follows:

i. Written and Oral feedback from focus groups (qualitative)

ii. Case study feedback produced as documents (qualitative/quantitative)

iii. Questionnaire responses from all participants (qualitative/quantitative)

iv. Feedback from one-on-one interviews (qualitative/quantitative)

v. Video footage of interviews for reviews (qualitative)

These variables will be core to our research but will help in developing important information for our dependent variables which were listed earlier (age, race, gender, religion, and ethnicity). Apart from these dependent variables there could be more dependencies that could be added based on the results of our data collection which will help increase the accuracy of the quantitative research for algorithm development that may follow this.

E. *Plan for Data Analyses*

Since this research is dealing with both qualitative and quantitative types of data, we will follow a mixed approach to analyze our data. The overall major focus of this research is to produce a dataset that categorizes al possible bias into different dependable categories. So that this dataset can be further used to train a new and improved Sentiment Analyses algorithm by adding the dependencies produced by this research. The feasibility of success will be analyzed by further research and is not part of this research.

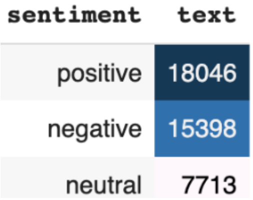
All data collected from focus groups that identifies a certain type of bias will be manually analyzed by our team at the end of each session to find repetitions in the data gathered that may represent a certain type of bias. Then such findings will be categorized into our dependable categories (age, gender, race, religion, and ethnicity). Following the same procedure data collected from case studies will first be analyzed manually and then findings will be categorized into categories.

An extensive content analyses will be done for all the answers generated from the questionnaire and based on the analyses; further dependencies may be added to the existing list. Interview footage will be peer reviewed to understand progress and create a classified dataset.

Since we are trying to build a very large dataset that includes both qualitative and quantitative data, this research will also analyze the methods that can be used to compress such data to increase the efficiency of computational analyses. We understand from [7] that it is much easier to encode qualitative data in numerical formats to process and understand data faster in a computational sense. We will use a simple encoding and decoding technique to encode all data into categories. Each value in each category will also carry the sentiment value along with it, i.e., positive, negative, or neutral. This will help in visualizing our data and ensure we have almost equal number of positive, negative, and neutral distinctions in our data which is essential to produce an accurate language processing model. Based on the results we can take further actions to improve our data or move forward with the research.

F. *Plan for Data Presentation*

Based on the nature of this research and learnings from the literature review [2] [4] [6] [7], it is only appropriate to use tables, charts, and figures to represent the findings. Once all data is collected and compiled through the data analyzation techniques discussed, the visual representation of the overall analyses would look like the example presented below:



sentiment	text
positive	18046
negative	15398
neutral	7713

Fig. 2. Sentiment table presentation

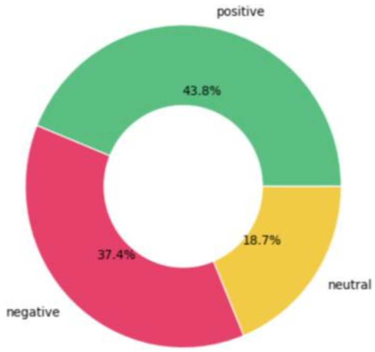


Fig. 3. Sentiment Pie Chart representation

Following these ideologies, we have categories as below:

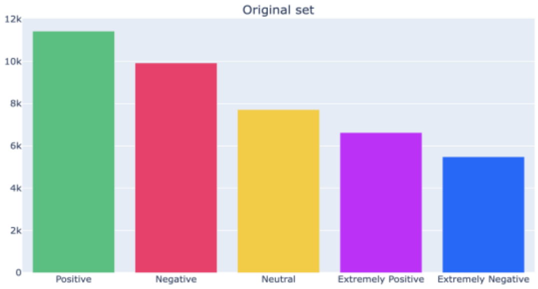


Fig. 4. Graphical representation of distinctive sentiment categories

Using these ideologies, we can further develop strategies to help visualize all other computational data as well, so that it can easily be captures by the reader and a common understanding can be formed.

G. *Evaluation and Validation Strategies Used*

This research has various quality assurance and ethical validation steps embedded into the sampling methods as well as the methods and instruments used for procuring data. Using a quality and performance check at multiple steps, this research will evaluate and validate the data set which will help create the basis for further research as that is our primary objective.

The added step of using a case study to build our dataset is solely done to ensure that the data collected is valid and can be further analyzed for its quality by having people in prison or individuals convicted by court to participate in this research. As they are convicted for racial (or other abuses) by law, we know that we can trust the information provided by them to validate our data.

It is also understood that there still may exist certain accepted stereotypes that might not trigger a negative bell for a certain group of people, but it may for another. To handle such bias is hard and there is no existing way understand to collect and compile such data.

III. Conclusion and Future Work

Mentioned earlier, this research will produce an extensive dataset that can be further used to train improve the existing Sentiment Analyses Algorithms. This research would help in building trust between the users and service providers as well as government entities who all aim to indulge users in their services for longer periods of time.

On successful execution and implementation of this research, we expect to find less hate speech on online platforms and more accurate recommendations based on an individual’s background and personal beliefs. This would promote more personalized content and improved user activity on all platforms. Thus, increasing the overall time spent online, helping online partners generate more revenue. Thus, it is essential to build and collect appropriate data via this research.

IV. Acknowledgements

I would also like to take this opportunity to acknowledge and thank the unit convener Mr. Dharmendra Sharma for providing this opportunity to work on an extensive research project. I would also like to thank Dr. Asmaa Elsaeidy for her help and guidance throughout this venture. It would not have been possible to achieve any results, without her continuous support. I deeply appreciate it.

References

[1] M. N. Al-Kabi, A. A. Al-Qwaqenah, A. H. Gigieh, K. Alsmearat, M. Al-Ayyoub and I. M. Alsmadi, "Building a standard dataset for Arabie sentiment analysis: Identifying potential annotation pitfalls," *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, Agadir, Morocco, 2016, pp. 1-6, doi: 10.1109/AICCSA.2016.7945822.

[2] T. Davidson, D. Bhattacharya, and I. Weber, “Racial Bias in Hate Speech and Abusive Language Detection Datasets,” *arXiv:1905.12516 [cs]*, May 2019, Available: https://arxiv.org/abs/1905.12516

[3] S. Kiritchenko and S. M. Mohammad, “Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems,” *arXiv.org*, 2018. https://arxiv.org/abs/1805.04508

[4] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, “Sentiment Analysis of Twitter Data,”Association for Computational Linguistics, 2011. Available: https://aclanthology.org/W11-0705.pdf

[5] Thelwall, M. (2018), "Gender bias in machine learning for sentiment analysis", *Online Information Review*, Vol. 42 No. 3, pp. 343-354. https://doi.org/10.1108/OIR-05-2017-0153

[6] M. Diaz, I. Johnson, A. Lazar, A. M. Piper, and D. Gergle, “Addressing Age-Related Bias in Sentiment Analysis,” *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Apr. 2018, doi: https://doi.org/10.1145/3173574.3173986.

[7] A. Severyn and A. Moschitti, “Twitter Sentiment Analysis with Deep Convolutional Neural Networks,” *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Aug. 2015, doi: https://doi.org/10.1145/2766462.2767830.