# Sentiment Rating for bias: An Independent Study

**Kausik Lakkaraju**
Doctoral Student
Artificial Intelligence Institute
University of South Carolina
kausik@email.sc.edu

## Abstract

Sentiment Analysis Systems (SASs) are known to exhibit gender and racial bias. In a previous work on sentiment rating [1], we proposed a rating system to evaluate these SASs to know how trustworthy they are. We also considered a composite scenario involving SAS and translators as translators are also prone to exhibit bias. In addition to what we have done before, I will discuss the possibility of performing similar experiments in a multi-modal environment.

## 1 Introduction

Extensive research is being done in the area of sentiment analysis, especially, on how data can be mined from various sources and pre-processed efficiently so that its sentiment can be estimated. This research is prominent in the field of healthcare. For these kind of experiments, existing SASs are being used but they are not being evaluated for any bias before performing the actual experiment as it will be hard to evaluate the trustworthiness of each system before using them. Sometimes, people might not even acknowledge that there is a possibility of such a bias in the system. In our previous work, we have explored different SASs, the gender bias that they exhibit and also proposed a rating system which would help us to measure how trustworthy a SAS can be.

In this study, I would like to focus more on how we can extend the existing work that we did to more diverse scenarios like applying similar techniques in a multi-modal scenario.

This report is organized as follows: My contribution in the paper that we worked on will be briefly discussed in the existing work section. The related work will be discussed in the literature survey section. The future work section describes how this work can be extended in future.

[1] This is a work yet to be published.

## 2 Existing work

My advisor, Dr. Biplav Srivastava, and I have worked on a paper in which we have explored the gender bias that SASs exhibit and we have also considered the composite case of translator and SAS. We have considered a dataset from (Kiritchenko and Mohammad, 2018). We split the whole dataset into biased and unbiased datasets. My part of the work was to collect different off-the-shelf sentiment analyzers along with neural network based sentiment analyzers and calculate sentiment scores, in the case of SAS and other results like word difference and gender difference when the sentences are translated from English to French and back to English. Several interesting results were observed. Refer to the experiments section and Behavior of translator and SAS together section of the paper to learn more. The rating method which would finally say whether the system is biased or not is developed by my advisor based on his previous work (Srivastava and Rossi, 2020). This was explained in 'Sentiment Rating System' section and also 'Rating Composite Services: SAS and Translator for Multi-Lingual SAS' section of the paper.

We have considered well-known SASs like *TextBlob* and *VADER* along with 3 neural network based SAS: *CNN*, *GRU* and, *LSTM*. Recently, I have also added a transformer based SAS, *DistilBERT*. The results can be observed from the 'Results' section of the paper.

From the next section, I will be discussing how we could extend the current work and the related work which would be helpful for us to reach our goal.

## 3 Literature survey

In the paper (Truong and Lauw, 2019), the authors proposed *'Visual Aspect Attention Network'* (VistaNet) which relies on visual information as alignment for pointing out the important sentences of a document using attention.This work is the first to incorporate images as attention for review-based sentiment analysis. In addition to use the multi-modal nature of such reviews, we would also like to rate the systems which perform multi-modal sentiment analysis. These systems can be called as Multi-modal Sentiment Analysis Systems (MSSA).

Another paper ,(Pena et al., 2020), talks about gender and racial bias that is observed in multi-modal AI. They proposed a new framework called FairCVtest which is an AI-based automated recruitment system to study how multi-modal machine learning is affected by biases present in the training data. They manually scored these synthetically made resumes with gender and racial bias. They have developed a learning method based on the elimination of such sensitive information in multi-modal approaches, and apply it to their automatic recruitment testbed for improving fairness. They have made the framework publicly available. This dataset could be used for our own experiments as it has around 24,000 resumes.

(Soleymani et al., 2017) talks about various interesting works related to multi-modal sentiment analysis. They have cited a lot of related work and discussed about these in their paper. This is a very interesting study.

One interesting work which was mentioned in the above mentioned paper is (Borth et al., 2013).They have proposed a 'Visual Sentiment Ontology' (VSO) constructed from content on the web consisting of more than 3000 adjective-noun pairs like crying baby and a Sentibank library which could identify any of the 1200 adjective-noun pairs present in an image. They observed that their system outperformed the text-only sentiment analyzers. Works like these can be considered for sentiment rating.

According to a recent article, (Gershgorn, 2021), GPT-3 exhibited bias towards Muslims. GPT-3 created sentences associating Muslims with shooting, bombs, murder, and violence.

## 4 Key Proposals

After referring to some works, I feel that there are many unexplored problems within this domain. Sentiment analysis in multi-modal environment itself is a less explored area. Rating of such systems would be a good problem to explore. In addition to what we have, the current work and, possible future work, I would also like to make some key proposals with this study which are worth exploring.

- Making a new dataset that would be useful for our experiments or modify the existing ones.
- Bias might be observed in other multi-modal systems like image captioning systems. This has to be studied.
- Some datasets with image-caption pairs are available. For this kind of data, we can estimate the sentiment scores for each of the image and text separately. We could also generate caption for that image and calculate its sentiment separately. These kinds of composite experiments would lead to some interesting results.
- Throughout the study, I have only looked at papers on visual-based systems or text-based systems. Another modality, speech, also needs to be considered. This can also be an interesting line of work.

## 5 Future work

The above mentioned key proposals are within the scope of the paper. Some other topics can also be explored as a separate yet related work. GPT-3 can also be explored for bias. In addition to SASs and translators, other AI systems like chatbots, image captioning systems and, speech-to-text converters can be studied for bias.

## References

Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232.

Dave Gershgorn. 2021. 'for some reason i'm covered in blood': Gpt-3 contains disturbing bias against

muslims. In *https://onezero.medium.com/for-some-reason-im-covered-in-blood-gpt-3-contains-disturbing-bias-against-muslims-693d275552bf*.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana, June. Association for Computational Linguistics.

Alejandro Pena, Ignacio Serna, Aythami Morales, and Julian Fierrez. 2020. Bias in multimodal ai: testbed for fair automatic recruitment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–29.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.

Biplav Srivastava and Francesca Rossi. 2020. Rating ai systems for bias to promote trustable applications. In *IBM Journal of Research and Development*.

Quoc-Tuan Truong and Hady W Lauw. 2019. Vistanet: Visual aspect attention network for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 305–312.