Wentworth Institute of Technology
COMP7800 – Natural Language Processing
Summer 2022, Folajimi

Research Assignment 1 (RA1)

1) Choose an option form option 1 or option 2 below and prepare a paper of 2-3 pages (1.5 line space 12 points Times New Roman)

Option 1

After the M.I.T. Media Lab researcher Joy Buolamwini discovered faults in facial recognition technology, she published "Coded Bias", a documentary that analyzes algorithm bias.

1) Watch the "Coded Bias" documentary on Netflix, or any alternative media to understand the implications of algorithmic bias in AI and Machine learning, and how this might impact research and development in data-sparse regions of the world. Of course, this is a very significant issue in natural language processing, as biased NLP algorithms have the potential for serious negative effects on society by discriminating against people with a limited NLP dataset.
2) Conduct an in-depth literature search to find the trend of bias in NLP resources, and how this can impact communities in date sparse regions. For example, in a region like Africa alone, it has been reported that there are over 2,000 living languages as of 2021 (https://www.statista.com/)
3) Discuss the efforts that have been made within the NLP research communities and organizations to address the challenges of the sparsity of corpus and other language resources to support minority languages
4) Based on your literature search discuss the trend, successes, challenges, and future plans for eradicating racial bias in NLP
5) What suggestions or recommendations would you propose, based on your findings?
6) Write a research paper (2-3 pages) to report your efforts in addressing points 1-4 above.
7) Your paper must include at least 5 research paper citations with appropriate references using a standard referencing style

Here are some interesting ideas that you can use in your paper. These are just examples and you don't have to use them. Be creative and use catchy subheadings.

- What are the implications of racial bias in NLP?
- Within the context of NLP and racial bias, express your optimism or apprehension regarding the future of African-spoken languages
- What are the challenges of racial bias, as regards technologies like facial recognition and NLP. How are these challenges being addressed?

Option 2

Download the following South African Disinformation [Fake News] Website Data – 2020 (https://zenodo.org/record/4682843#.YfF2yerMJD8)

- Fake News (hinnews.com) Questionable.xlsx
- Fake News (mzanzi stories) Fake.xlsx
- Fake News (sa-news.com) Fake.xlsx
- Fake News (search67.com) Fake.xlsx
- Fake News (whatsappgroup.co.za) Fake.xlsx

You are required to extract some useful insights from the downloaded data by using some relevant NLP techniques, Python libraries, or other useful resources, and write a paper discussing some interesting findings. Examples of such analytics might include but are not limited to:

- Check whether there are some common words occurring in each category of fake news
- Find out which category of fake news is most popular, or has the most content
- Which word has the highest occurrence in each category? Is there a relationship between the word's occurrence and its occurrence in other categories?

Your paper should include a section discussing at least 5 related literature in fake news, with their citations and references using a standard referencing style.