Data Mining Human Reasoning

*Mid Term Report*

*Project Author: Taha Kashaf*

*Advisor: John Lawrence*

**Aims and Objectives:**

In today’s world, a controversial talking point is one of vaccines. This is of course very topical right now, in the midst’s of the worlds best documented pandemic: COVID-19. However, while many of us see vaccinations and large-scale inoculation as the final solution to bring an end to 3 years of measures and restrictions, many also see vaccinations as a bad thing.

Now, anti-vaccination rhetoric has been around long before COVID-19, however the visibility of the movement has grown dramatically, and its effects on the rest of us have become more tangible. As we see reports of hospital urgent care wards filled primarily with unvaccinated covid patients. (<https://northeastlondonccg.nhs.uk/news/almost-90-of-patients-admitted-to-intensive-care-units-in-north-east-london-are-not-fully-vaccinated/>)

So, the problem here is an obvious one: large groups of people carry a negative sentiment towards vaccinations. The goal of this project is simple. To use a data-driven approach, powered by machine learning and sentiment analysis to attempt to understand *why* people have the opinions they do regarding vaccines.

This is obviously not a solution to the problem, but with a topic as complex as human reasoning, and why people choose the things they do. First, you need a comprehensive understanding of why. Once that understanding has been built, only then can you begin to tackle the problem. This project is an attempt at understanding the reasoning behind the problem.

**Background Research & Takeaways**

The area of big data is a rapidly growing one – and one that has come to the forefront of discussion in recent times. Recent large-scale political campaigns have been won and lost on data-centric approaches. (<https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie>) The success of these campaigns has even lead to legislation being passed in the area, such as GDPR (the General Data Protection Legislation).

So, there is no question about the importance or research potential within big data, the ability to analyze and extract value from large scale unsorted data is an ever-growing field. One of the many ways in which this is possible is sentiment analysis. Once again, an ever-growing field.

There are 3 main types of sentiment analysis: (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7827575/>)

* Lexicon-based approaches.
  + These rely on corpora or dictionaries that contain terms classified by their sentiment.
* Machine-Learning approaches.
  + Supervised learning techniques
    - These use algorithms trained by pre-labeled data
  + Unsupervised
  + Semi-Supervised
    - A combination of supervised and semi-supervised, in an attempt to design the most appropriate classification model.
    - A hybrid of lexicon-based, and machine-learning based.

In recent times, similar studies have been carried out, using social media and in particular twitter data to analyze vaccine sentiment. Much as I plan to do myself. However, these studies focus primarily on classifying the data and presenting the empirical evidence as the results.

The goal of my own project is both to classify the data as a first step, but then to look beyond that and use the results to search for further correlations that will aid in understanding *why* the sentiment is held in the first place.

Lastly, the final takeaway I have from the research and similar studies I have found so far, is that they are all primarily presented in strictly academic ways. At best, the least academic iterations are articles that regurgitate the summaries provided by the studies. Due to this, I’ve decided that I want to present my findings in a way that is easily and quickly understood. Something that does not require reading large paragraphs populated by statistics to understand my conclusions.

Due to this, the final part of my project is currently planned to be a website where the results are presented in a user-friendly and easily understood way. Attention is a limited and high competed for resource, so I wish to make the best use of any attention my project may receive, providing conclusions for those who just want to look over it quickly, while still retaining the detail for those who want it.

**Main Features**

The main feature the project aims to achieve is a website that presents the conclusions of my research in a clear and user-friendly way, as described above. This is the only “application” that I aim to develop, and how my project will be interacted with by any users.

Of course, that is purely front-end. Most of the work on the project will be work that won’t be seen to the final user as it will exist on the back end. This involves