TODO

Henry Ash Williams

May 16, 2024

Contents

1 Introduction **1**

2 Background Research **1**
2.1 Data Analysis 1

3 Methods **2**

4 Evaluation **2**

5 Conclusion & Further Work **2**

1 Introduction

Propaganda is a technique in which one group attempts to manipulate the opinions, beliefs and attitudes of another. This technique has been utilized extensively throughout history and is even seen in contemporary media. For example, shortly before the death of Augustus, the first emperor of Rome, an inscription was made on what was to be his final resting place. The inscription is known as "The Deeds of the Divine Augustus", and outlines both his religious power by describing himself as 'Divine', and his importance to the Pax Romana, or the Roman Peace, by outlining his strategic ability through the anecdotes of his conquests found within the inscription. This was an attempt to shape the attitudes of his citizens towards him to be more favorable. In the years following his death, he was remembered as "The Divine Augustus". However, during his reign, he routinely suppressed dissidents of the Roman Empire in an attempt to centralize his power. While it cannot be told whether the favorable view of Augustus within the Roman Empire was solely due to this text, this was not the only piece of propaganda he produced, and it likely helped to shape the view of his citizens, without them even noticing that their opinion was manufactured.

This technique is not restricted to ancient history and is often seen in our modern media, whether we notice it or not. For example, in the lead-up to and during America's invasion of Iraq, US media reported on an alliance between Saddam Hussein and al-Qaeda in order to justify a US-led invasion of Iraq. The two groups were, however, largely at odds with each other, and no such allegiance was ever proven. Despite this, in 2002 a majority of Americans believed that Saddam Hussein was directly involved with both al-Qaeda and the 9/11 terror attacks.

A majority of people who are exposed to propaganda do not recognize it as such, and considering the impacts of believing propaganda, both from a security perspective and a social justice perspective, can lead to significant consequences. However, by nature, propaganda is extremely difficult to detect, and many people struggle to recognize it even today. Considering the amount of media produced and consumed each day, a human approach to propaganda detection and classification is impossible, and a new approach is required.

This paper will explore the effectiveness of modern machine learning models in both these tasks, namely the detection of propaganda within text, and the classification. Classification in this case refers to the methods of propaganda used within the text. This includes appealing to an individual's sense of fear, exaggerating facts, or the oversimplification of a complex topic. In this paper, we will cover my approaches to these tasks, using natural language processing techniques such as Term Frequency-Inverse Document Frequency, and Unigram Precision, as well as more complex, large language model approaches such as the fine-tuning of BERT models.

2 Background Research

2.1 Data Analysis

To develop a model capable of achieving the goals outlined in the previous section, we need a thorough understanding of the data we've been provided with. The dataset of text samples, some of which contained propaganda, and others that didn't. Overall, there were 3,200 samples, 50% of which contained propaganda, and 50% of which didn't. It came in the form of two separate tab-separated value files, one for training, and another for testing. The testing dataset contains 640 samples, 48% of which contain propaganda. The training dataset contains 2,560 samples, 52% of which contained propaganda.

There are 9 labels, including the label for samples without propaganda. The average word count of each class is displayed in Figure 1

Within each of the text samples, there are tags indi-

Label Name	Average Sample Length
Appeal To Fear Prejudice	39.6
Name Calling, Labeling	44.5
Exaggeration, Minimisation	40.0
Not Propaganda	30.0
Loaded Language	37.7
Causal Oversimplification	43.6
Doubt	41.6
Repetition	34.4
Flag Waving	40.0

Figure 1: The name of each label followed by its average word count

cating the area of interest. This indicates where the propaganda is within the sample. Upon extracting the text within this area of interest, the

3 Methods

4 Evaluation

5 Conclusion & Further Work