A Machine Learning Approach for Detecting Propaganda in News Articles with Case Study on Israel-Palestine Coverage

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Artificial Intelligence and Ethics

Student: Reuben Lesser
1st Supervisor: Associate Professor Dr. Mahsa Abazari
Co-Supervisor: Assistant Professor Dr. Adrian Hillman

Northeastern University London

Word Count: 9,870

Abstract

Propaganda detection in news media remains a critical challenge as misinformation continues to proliferate online. This paper investigates the efficacy of machine learning approaches in automatically identifying and classifying propaganda in news articles. Two models were developed and compared: a Convolutional Neural Network (CNN) and a Conditional Random Field (CRF), both utilising BERT tokenisation. The final model was designed to perform a two-step classification process: first identifying sentences containing propaganda, then labeling specific propaganda techniques within those sentences. Results demonstrate the feasibility of automated propaganda detection in news articles. Deployment on articles discussing the current events in Israel/Palestine show no significant difference in propaganda prevalence for articles relevant to the conflict versus irrelevant articles. While initial metrics indicate a promising implementation of this technology, further refinement is necessary before real-world deployment. This work contributes to the growing field of automated content analysis and provides a foundation for developing more sophisticated propaganda detection systems that enhance media literacy, and empower readers to better understand and critically evaluate the news they consume.

Acknowledgements

I would like to thank my advisors, Dr. Mahsa Abazari Kia and Dr. Adrian Hillman, for their guidance throughout this process, my friends who generously helped review and edit this dissertation, and my fellow graduate students whose support and friendship made this journey both possible and memorable.

Contents

1. Introduction

In an era characterised by information overload and the rapid dissemination of news through digital platforms, the ability to discern factual reporting from erroneous or biased information has become increasingly important. As individuals, we are constantly bombarded with a vast array of news sources, each with its own agenda and level of accuracy. This wealth of information can make it challenging to separate truth from fiction, leading to the spread of misinformation, disinformation, fake news, and propaganda [1].

Distinguishing between factual reporting and propaganda has become essential for maintaining an informed citizenry and fostering a healthy democratic discourse. For instance, during the Israel-Hamas conflict that began in October 2023, a misleading video circulated on social media claimed to show Palestinians faking a funeral, gaining traction among users and even being shared by prominent figures. However, fact-checkers quickly revealed that the video was actually an unrelated clip from 2020, highlighting the critical need for media literacy and careful verification of information during times of conflict when misinformation can easily spread and influence public perception [2]. The task, however, of independently verifying the accuracy and reliability of every piece of information we encounter is almost insurmountable, as the sheer volume of content combined with the sophisticated techniques used to spread misinformation can be overwhelming [3]. Furthermore, our own biases and preconceptions can inadvertently influence our judgement, making it difficult to approach information objectively.

Given these obstacles, it is clear that developing effective tools and strategies for information evaluation is vital. Previous research in automated propaganda detection has largely focused on two main approaches: document-level classification and sentence-level analysis. For example, one study using document-level classification created a corpus in which entire news articles were labelled in four categories (propaganda, trusted, hoax, or satire), using distant supervision by assuming all articles from a given source shared the same label [4]. While this provided a base for automated detection, this broad classification at document-level and based on article source failed to capture the nuanced ways propaganda can manifest itself, as it failed to consistently capture sudden spikes in public opinion.

Another example of automated propaganda detection is seen in *Proppy*, one of the first real-world, real-time, fine-grained propaganda detection systems designed for online news [5]. This system further demonstrated the feasibility of automated propaganda detection and suggested models based on character versus word n-grams (sequences of n consecutive characters or words),, showing that increased granularity may yield better results.

In 2020, the International Workshop on Semantic Evaluation announced a task based on automated propaganda detection. They invited competitors to create models to classify news articles as propaganda both on the sentence and technique level, with the former classifying full sentences as propaganda or non-propaganda, and the latter identifying spans of text within sentences that contain propaganda and labeling them with one of 14 propaganda techniques [6].

This paper aims to build upon the models developed during said task by utilising the same dataset and strategies based on the top performers of this task. Beyond automated analysis, it also aims to enhance digital literacy by helping users understand common propaganda techniques, recognise coordination patterns in information campaigns, and develop independent critical thinking skills. This will be done by deploying the model on a set of articles from the BBC and *The Guardian* discussing the current Israel-Palestine conflict.

The proliferation of propaganda is a phenomenon that has become particularly evident in coverage of the events occurring in Israel and Palestine since October 7th, 2023 [7]. This complex information environment presents an ideal case study for testing and validating these propaganda detection models, as it encompasses multiple forms of media manipulation, and provides an opportunity to evaluate the tool's effectiveness and practical utility in real-world scenarios. The analysis will focus on identifying potential biases, misleading narratives, and propaganda techniques employed in the coverage of the Israel-Palestine conflict. By empowering individuals to critically assess news content related to this conflict, the tool can contribute to a more nuanced understanding of the situation. This in turn can help readers form opinions that are based in fact, rather than stereotypes of misleading narratives perpetuated by biased reporting and propaganda.

Note that while the author holds personal views on this topic, this paper aims to present an impartial analysis based on factual information and diverse perspectives.

2. Background

2.1 Defining False Information and Propaganda

False information can present itself in a number of ways. Three terms most commonly used to describe types of false information are misinformation, disinformation, and fake news. Misinformation can be defined as “constituting a claim that contradicts or distorts common understandings of verifiable facts” [8]. It is unintentional in its dissemination of false or misleading information. Contrast this to disinformation, which is regarded as the subset of misinformation that is deliberately propagated. Then there is fake news, or deliberately misleading articles designed to mimic the look of actual articles from established news organisations [8].

Propaganda, however, represents a similar but distinct phenomenon with its own theoretical foundations. Encyclopedia Britannica defines propaganda as “dissemination of information—facts, arguments, rumours, half-truths, or lies—to influence public opinion.” [9]. Tucker et al. define it as “information that can be true but is used to ‘disparage opposing viewpoints’.”[10], while Zannetou et al. argues propaganda is a special instance of fabricated stories that aim to harm the interests of a particular party and usually has a political context [11]. Despite their differences, these definitions all focus on the deliberate manipulation of information to shape political beliefs and public opinion, highlighting how propaganda serves as a tool for influencing societal attitudes towards specific issues or ideologies. Unlike misinformation and disinformation, which are primarily defined by their relationship to factual accuracy, propaganda is distinguished by its persuasive intent and strategic deployment of communication techniques.

2.2 History of Propaganda

Systematic study largely focuses on its use in 20th-century conflicts, particularly during the World Wars and the Cold War period [1], [12], but propaganda has existed since early civilisations [13. In ancient civilisations, propaganda relied heavily on visual imagery and oral traditions. The Neo-Assyrian Empire (911-609 BC) exemplified this through palace art depicting divinely sanctioned military campaigns [14]. Ancient Greece (1200-323 BC) also exhibited sophisticated propaganda, with the rise of democracy in Athens leading to new forms of public persuasion. Athenian orators developed methods of presenting historical events in ways that served their political aims by selectively emphasising certain aspects of history while downplaying others, creating compelling narratives that connected

past events to present political decisions [15]. During the Roman Empire, Emperor Augustus wrote the autobiographical Res Gestae Divi Augusti and strategically distributed it throughout the empire, inscribing it on public buildings and monuments to ensure his carefully crafted message reached the widest possible audience [16].

During the Middle Ages (476-1450 AD), religious institutions, particularly the Catholic Church, became major propagandists. The Crusades (1095 – 1291 AD) marked a pivotal moment in propaganda history, with the Church employing multiple channels such as architectural symbolism in cathedrals, pilgrimage routes, and public ceremonies to reinforce its authority [17]. The invention of the printing press in the 15th century revolutionised propaganda dissemination, with Martin Luther's Protestant Reformation (1517) showcasing its potential for mass persuasion through the German Bible translations and provocative woodcut illustrations [18]. This period also saw the emergence of political cartoons and pamphlets as powerful propaganda tools, particularly during the American and French Revolutions, such as Thomas Paine's 'Common Sense' (1776) which exemplifies how printed propaganda could shape public opinion [19].

The 19th and 20th centuries witnessed a dramatic increase in propaganda's reach and sophistication due to mass media, increased literacy, and new technologies such as radio and film. World War I marked arguably the first large-scale, organised propaganda campaigns by governments, with dedicated propaganda offices established in many countries [20]. The interwar period and World War II saw further refinement of propaganda techniques, especially by totalitarian regimes. Nazi Germany's Ministry of Public Enlightenment and Propaganda demonstrated how modern media could be systematically employed for mass manipulation [21]. The Cold War era (1947-1989) was characterised by ideological warfare between capitalist and communist blocs, using media, culture, and education for propaganda purposes. The Soviet Union's Agitprop system and the United States Information Agency engaged in sophisticated campaigns, using everything from jazz music to space achievements as propaganda tools [22].

Today, the internet and social media have revolutionised propaganda dissemination, allowing for targeted campaigns and the rapid spread of disinformation. Artificial intelligence and machine learning algorithms also enable even more highly sophisticated campaigns, with deepfake technology and automated content generation presenting new challenges for information integrity [23]. A prime example is the Russian 'Doppelganger' operation, first established in 2022, which creates sophisticated fake news websites and AI-generated content to spread disinformation across multiple platforms, targeting specific audiences with divisive messaging [24]. With a large proportion of propaganda existing online these days and the advanced nature of its generation, detection and mitigation have become increasingly complex, requiring innovative

approaches and technologies to preserve the integrity of public discourse and democratic processes.

2.3 Propaganda in Israel-Palestine News Coverage

The current news coverage of the Israel-Palestine conflict appears to be deeply saturated with propaganda from various sides, making it challenging for audiences to discern objective truth [7]. There have been incidents of debunked information circulating during the conflict, such as exaggerated claims regarding atrocities and shifting narratives surrounding specific attacks, which highlight the challenges of verifying facts in a rapidly evolving media landscape [7].

Major Western publications and media outlets have shown a consistent bias towards Israeli narratives, often employing emotive language to describe Israeli casualties while presenting Palestinian deaths in more abstract terms [25]. This manifests through selective emphasis on certain events, careful choice of terminology, and strategic use of visual imagery [25]. Palestinian and Hamas-affiliated media frequently engage in antisemitic rhetoric and imagery, promoting harmful stereotypes and conspiracy theories about Jews rather than focusing on substantive critiques of Israeli government policies [26]. This approach not only undermines legitimate Palestinian grievances but also fuels the cycle of mistrust and hatred.

The emergence of new content manipulation technology has further complicated the media landscape, making verification increasingly challenging. Coordinated information campaigns, both organic and automated, can rapidly spread misinformation across multiple platforms before fact-checkers can respond [27]. This has led to a polarised media landscape, making it increasingly difficult for the public to form a comprehensive and balanced understanding of the issues at hand.

These factors make this an ideal case study for testing the utility of a propaganda detection and information tool in navigating news coverage. By analysing reports from the BBC and *The Guardian*, it could help readers identify potential biases in how Israeli and Palestinian perspectives are presented, particularly focusing on the nuanced ways that language choices and framing devices shape narrative construction. The reputation of these sources largely removes the intent to create deliberate propaganda and disinformation, allowing us to focus more specifically on misinformation and unintentional bias. Both outlets are widely recognised for their journalistic integrity and commitment to balanced reporting [28]. The BBC, in particular, consistently ranks as one of the most trusted news sources in the UK, with 44% of Britons considering it “very trustworthy” or

“trustworthy”. Similarly, *The Guardian* is highly regarded, especially among left-leaning audiences, with a net trust score of +15 [28].

2.4 Propaganda Detection Overview

The systematic study and detection of propaganda presents several unique challenges that require advanced technological solutions. From varied definitions and historical analysis, propaganda can contain both factual and inaccurate information, often employing rhetorical techniques and psychological manipulation that leverage powerful emotions like fear and anger to sway public opinion [8]. Additionally, the subtle nature of modern propaganda techniques makes it difficult to distinguish between biased reporting and factual news coverage without sophisticated analysis [6].

These challenges have been recognised for nearly a century, with early efforts to combat propaganda's influence dating back to the pre-World War II era. The foundations of systematic propaganda detection were established by the Institute for Propaganda Analysis (IPA), which operated from 1937 to 1942 and “advocated for ‘propaganda literacy’ against the backdrop of rising nationalism before and during World War II” [30]. The IPA developed pioneering frameworks including their ‘ABCs of Propaganda Analysis’, which encouraged readers to identify conflicts and question the independence of their opinions. Their work included publishing monthly bulletins and leaflets analysing Nazi and fascist propaganda, and addressing topics such as the importance of critical thinking in education and journalism [30].

Moving beyond the IPA’s public education focus, the Cold War era marked a significant evolution in propaganda detection methods as they became more sophisticated and institutionalised. President Ronald Reagan's administration in the 1980s set up the Active Measures Working Group which developed developed systematic approaches to collect, analyse, and publicise suspected Soviet disinformation [22]. Western intelligence agencies built networks of informants and employed advanced surveillance techniques to track disinformation campaigns, while media outlets like Radio Free Europe dedicated resources to exposing Communist misinformation. As technology advanced, intelligence agencies began incorporating computer analysis to identify patterns in Soviet disinformation campaigns, establishing the foundation for modern computational approaches [22].

Not until the late 2010s and early 2020s did computational models for propaganda detection gain significant traction. 2019 saw the introduction of *Proppy*, one of the first real-world, real-time monitoring systems designed to unmask propagandistic articles in online news [4]. This model used n-grams to demonstrate the feasibility of automated

propaganda detection at a high level. Then came neural architectures, for which one study compared Long Short-Term Memory (LSTM), hierarchical bidirectional LSTM (H-LSTM), and Convolutional Neural Network (CNN) models for classifying text as propaganda or non-propaganda [31].

The introduction of transformer-based models marked another significant development in the field. One study explored the use of Bidirectional Encoder Representations from Transformers (BERT), a neural network-based technique, particularly in fine-grained propaganda detection [32]. Their research focused on detecting both propagandistic sentences and specific techniques, demonstrating the potential of transformer models for this task. This research performed a fine-grained analysis of texts by detecting all fragments that contain propaganda techniques as well as their type, and created a corpus of news articles manually annotated at the fragment level with eighteen propaganda techniques. This study was later built upon during the 2020 International Workshop on Semantic Evaluation, in which teams were tasked with creating propaganda detection models at the sentence and span level using this dataset [5]. This task produced numerous results, providing a wealth of data for comparison among different approaches.

Despite these significant advancements in the field, there is still not a widely used, user-friendly propaganda detection tool available to the general public [33]. Many existing solutions are primarily geared toward researchers and professionals, requiring technical expertise that most users may not possess. As awareness of misinformation continues to rise, there is a growing demand for accessible tools that can empower individuals to critically evaluate the information they encounter.

2.5 Machine Learning

2.5.1 Overview

Machine learning (ML) is a subfield of artificial intelligence (AI) that enables computer systems to learn from data rather than following explicit programming rules [34]. This approach enables handling complex tasks like natural language processing and predictive analytics that traditional programming cannot effectively address.

The core functionality of ML comes from its ability to identify patterns in large datasets and apply these learned patterns to new, unseen data. These systems involve several interconnected components: the model architecture that defines how information is processed; training procedures that determine how the model learns from experience; evaluation metrics that measure performance; and critically, the characteristics and quality of the training data itself [35]. Through iterative optimisation of these components,

machine learning systems can achieve increasingly accurate results on their designated tasks.

2.5.2 Model Architecture

Model architecture in ML refers to the overall structure and organisation of components that define how a model processes information from input to output [36]. The choice of architecture is one of the most crucial decisions in ML design, as it directly impacts what kinds of patterns the model can learn and how efficiently it can learn them.

The two models used in this paper, Convolutional Neural Networks (CNNs) and Conditional Random Fields (CRFs), represent fundamentally different approaches to pattern recognition in ML. CNNs are deep learning models that process data through layers of learnable filters and feature maps, particularly excelling at tasks involving spatial hierarchies [36]. The architecture uses sliding filters across input to detect simple features in early layers and increasingly complex patterns in deeper layers, with pooling layers reducing spatial dimensions while preserving important features, and fully connected layers combining high-level features for final classification.

CRFs, on the other hand, are probabilistic graphical models that excel at structured prediction tasks where context and relationships between outputs are crucial [37]. Unlike CNNs, which process inputs independently through a feed-forward architecture, CRFs explicitly model dependencies between predictions. Their architecture consists of nodes representing variables and edges representing dependencies, with the model learning potential functions that capture both local features and transition dynamics between states. The ability to model sequential or spatial dependencies makes CRFs especially powerful when combined with other architectures - for instance, many modern systems use CNNs to extract features and CRFs to ensure prediction consistency, combining the hierarchical feature learning of CNNs with the structured prediction capabilities of CRFs.

2.5.3 Training Algorithms

ML training algorithms are the systematic procedures used to optimise a model's parameters based on training data [35]. The optimisation process typically involves computing gradients that indicate how each parameter should change to minimise a loss function, which quantifies the difference between predicted and actual outputs - the smaller the difference, the better the model's performance. Common approaches include stochastic gradient descent, which processes small random batches of data to efficiently approximate the optimal direction of parameter updates; and adaptive learning rate

methods like Adam and RMSprop, which dynamically adjust how quickly parameters change based on past gradient information [38].

Each algorithm makes different trade-offs between computational efficiency, memory usage, and optimisation effectiveness. For example, the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm is a quasi-Newton method that approximates the Broyden-Fletcher-Goldfarb-Shanno algorithm using a limited amount of computer memory. This allows it to efficiently store and use information about recent gradients to estimate the function's curvature, allowing it to take intelligent steps toward optimal solutions while keeping memory usage low [39].

2.5.4 Evaluation Metrics

ML evaluation metrics provide standardised ways to assess model performance and compare different approaches across various tasks. These metrics quantify different aspects of model behaviour and accuracy and depend on the type of task being performed [40]. For example, classification tasks, such as the one performed in this paper, involve predicting discrete categories or labels. The model's job is to assign input data to a predefined class or multiple classes.

Classification metrics provide a framework for evaluating predictions across different categories. When assessing these predictions, correct and incorrect classifications are tracked in several ways. True Positives (TP) occur when a class is correctly predicted - for instance, when a sentence is predicted as propaganda and it is in fact propaganda. True Negatives (TN) represent correct predictions of other classes, such as when a text is predicted as non-propaganda and it is correctly non-propaganda. False Positives (FP) are when a class is incorrectly predicted - for example, a text is labelled as propaganda when it is not. Then False Negatives (FN) are when the correct class is not identified, so a text is labelled as non-propaganda when it is in fact propaganda [41].

These basic measures form the foundation for more complex metrics. Classification tasks employ metrics like accuracy, precision, recall, and F1-score, each highlighting different aspects of classification performance. Accuracy measures overall correctness, precision focuses on FP rates, recall examines FN, and F1-score balances both precision and recall. These metrics can be used for binary and multi-class scenarios. The principles remain the same either way as the measurement is always based on how well the predictions match reality [41].

2.5.5 Natural Language Processing

Natural Language Processing (NLP) builds upon the fundamental ML principles discussed above by teaching machines to understand and work with human language data. While traditional ML might work with structured numerical data, NLP tackles the complexity of text and speech by transforming linguistic information into formats that algorithms can process. This transformation can occur through several processes. Tokenisation breaks text into meaningful units (words or subwords) and is a necessary part of text processing. Other processes are not required but can enhance a model's success when used properly. Things like part-of-speech tagging which identifies grammatical roles, lemmatisation which reduces words to their base forms, and named entity recognition which identifies and classifies proper nouns, can enhance the model's understanding of linguistic structure and meaning. These steps enable more sophisticated analysis like syntactic parsing, which maps grammatical relationships between words, and semantic analysis, which extracts meaning from text [41].

NLP models learn patterns in language through various approaches, from statistical methods to neural networks, with different architectures specialised for specific tasks. These models learn to handle various challenges unique to language processing, such as ambiguity in word meanings, variation in sentence structure, and the contextual nature of human communication. This makes them particularly powerful for propaganda detection tasks. Advanced NLP techniques can identify subtle propaganda techniques by analysing rhetorical devices, emotional language, and logical fallacies, while also examining broader discourse patterns that might indicate manipulation attempts. This multilevel analysis, from word choice to overall narrative structure, enables systems to detect various propaganda strategies from simple name-calling to complex logical fallacies and emotional manipulation techniques.

2.6 Existing Work in Automated Propaganda Detection

Recent advancements in ML and NLP have enabled more sophisticated analysis methods. Large Language Models (LLMs) like GPT-3 and BERT have demonstrated significant potential in detecting propaganda in news articles [42]. These models can be fine-tuned on propaganda detection datasets, achieving high accuracy in identifying various propaganda techniques. Another promising approach utilises deep learning techniques, specifically CNNs and LSTM networks to classify news articles as propagandistic or non-propagandistic [43].

Several researchers have developed multi-task learning approaches that simultaneously identify propaganda techniques at both sentence and fragment levels [44].

While these systems achieve strong results, they typically require extensive computational resources and carefully annotated training data. Related work has explored attention mechanisms that can identify different propaganda techniques while providing interpretable results, though these approaches sometimes struggle with subtle forms of propaganda that require deep contextual understanding.

Hybrid systems combining rule-based and ML approaches have shown promise in providing explainable results [45]. These systems often incorporate predefined rules based on known propaganda techniques with flexible ML models that can adapt to new patterns. While this approach offers valuable transparency, especially for educational purposes, the reliance on predefined rules can limit adaptability to novel propaganda techniques.

Multi-lingual propaganda detection has emerged as an important research direction, using multilingual transformers to identify propaganda across different languages [46]. These approaches show promise in detecting propaganda that might otherwise escape monolingual systems, though performance varies significantly across languages and cultural contexts. Some researchers have incorporated domain adaptation techniques to improve performance on low-resource languages, but challenges remain in handling culturally-specific propaganda techniques.

Another significant advancement is seen in multi-modal approaches that analyse both text and images [47]. Systems that combine visual and textual analysis can identify manipulated content and misleading image-text combinations. While these approaches show promise in catching sophisticated propaganda techniques, they require significant computational resources and can produce false positives when dealing with legitimate artistic or satirical content.

As discussed in earlier sections, this paper aims to build off models developed as part of the SemEval-2020 Task 11 in the 2020 International Workshop on Semantic Evaluation [7]. The task attracted significant interest from the research community, with 250 teams initially registering to participate. Of these, 44 teams made official submissions on the test set, generating a total of 66 submissions across both subtasks. 32 of the participating teams also submitted system description papers detailing their approaches. The organisers were able to compile comprehensive information about the different modeling approaches through these system description papers. The techniques used by teams were catalogued in detailed comparison tables that broke down the key components of each system, such as the types of transformer models used, learning models employed, types of representations and features utilised, and additional techniques like ensembling and post-processing. This systematic documentation of approaches provided valuable insights into what modeling strategies were most effective for propaganda detection tasks.

Upon analysis of the top performing models, CNNs and CRFs consistently emerged as the most successful architectures, along with BERT tokenisation. BERT word embeddings were also used with significant success. Therefore, this paper aims to build upon these techniques to create and demonstrate a model with emphasis on a user-centric approach, real-time analysis, and critical evaluation of news content. Unlike existing systems that focus purely on automated detection, this approach aims to actively engage readers in the evaluation process while providing unbiased feedback to help mitigate confirmation bias. The system is designed to be accessible and user-friendly facilitating critical thinking rather than simply providing classifications of content. This approach recognises that effective propaganda detection must go beyond mere identification to foster an environment where users are open to examining their own biases and potentially revising their views.

A case study analysing media coverage of the Israel-Palestine conflict demonstrates the system's practical application in evaluating complex, polarising topics. This analysis showcases how the tool can be used to conduct larger-scale media evaluations, revealing patterns in coverage across different outlets and helping readers understand how various propaganda techniques may influence their perception of contentious issues. By combining technological solutions with psychological insights about bias and belief formation, this approach represents a step toward more nuanced and effective propaganda detection that prioritises user engagement and critical thinking.

3. Work Undertaken

3.1 Data Preprocessing and Analysis

3.1.1 Dataset

The dataset used in this paper was sourced from the SemEval-2020 Task 11 corpus, which is comprised of gold-level data of news articles annotated for propaganda techniques at the sentence and fragment level. The dataset contains 536 articles from 13 propaganda and 36 non-propaganda news media outlets, as labelled by Media Bias/Fact Check, a comprehensive resource that evaluates the bias and credibility of over 8,800 media sources, journalists, and politicians [48]. The articles were deduplicated on the basis of word n-grams matching and discarded faulty entries (e.g. empty entries from blocking websites).

The corpus underwent rigorous annotation by professional annotators from A Data Pro, a data solutions company that offers data labeling services, among others, to provide accurate and high-quality annotated datasets for various applications [59]. The annotation process involved two phases: first, two annotators independently labelled an article, then they collaborated with a consolidator to resolve any discrepancies in their annotations. The articles were annotated first at the sentence, with each sentence in a news article being marked as ‘propaganda’ or ‘non-propaganda’. Then text fragments were labelled with one of 14 propaganda techniques and marked by its starting and ending character positions within the text. The γ agreement score between annotators averaged 0.6, indicating reasonable inter-annotator agreement [7].

The fragment level dataset contains a significant class imbalance, with certain propaganda techniques appearing much more frequently than others. For instance, *Loaded Language* was the most prevalent technique and occurred about twice as often as the second most frequent technique *Name Calling*, while techniques like *Bandwagon* and *Reductio ad hitlerum* were relatively rare with only about 70 instances each. Originally there were 18 propaganda techniques, but rare techniques were merged based on similarity to ensure sufficient sample sizes for training. The technique *Repetition* was removed for this paper, as this would require more extensive implementation to track occurrences of text within an article, which is outside of the scope. This left 13 techniques for analysis.

The class imbalance in both datasets is addressed through upsampling, a technique that increases the representation of the minority class to match the size of the majority class. In this process, the minority classes are effectively duplicated until their size equals that of the majority class. This ensures that all classes are equally represented during training, helping to prevent bias towards the majority class.

An overview of all techniques used in this analysis is included in the appendix. See overviews of the dataset in Figures 1 and 2.

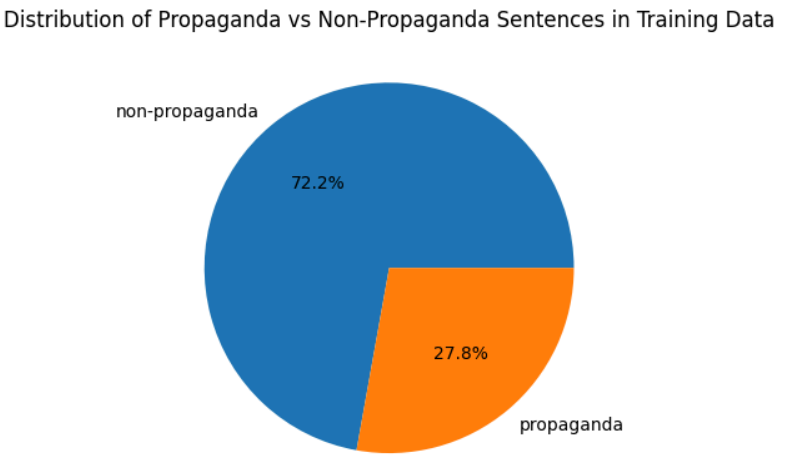

Fig 1.

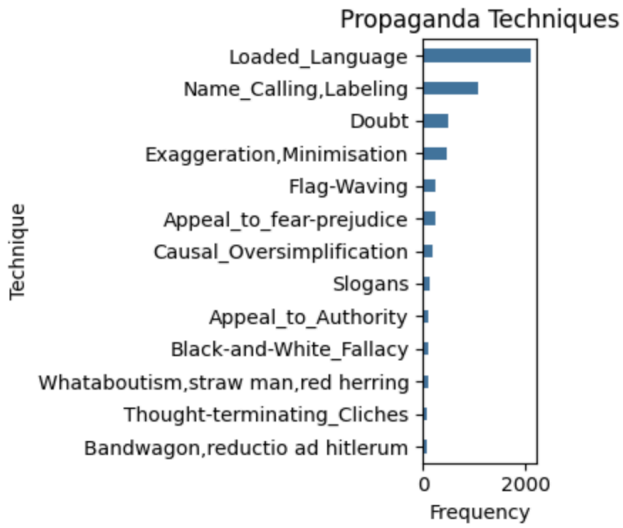

Fig.2

3.1.2 Text Cleaning and Normalisation

The text cleaning process implements a dual-path approach supporting both BERT and NLTK tokenisation strategies. This was originally done to compare the performance using each tokeniser, but BERT consistently showed better results so was used in the final model development and deployment. For text normalisation, stop words were removed using NLTK's English stop words list, and alphanumeric tokens were filtered. Other standard steps such as lowercasing and punctuation removal were initially performed, but did not yield significant results so were eliminated to reduce computational cost. Additionally, capitalised text and intentional punctuation use can indicate emphasis, emotional manipulation, or specific rhetorical techniques, so may be useful in propaganda detection. For the fragment-level implementation, text was split at white space and stemmed as this is standard practice when using the NLTK tokeniser.

3.2 Tokenisation

Two tokenisers are used in the final CRF model implementation: NLTK and BERT. The sentence level model uses BERT's pre-trained tokeniser (bert-base-cased) to process

the input text. While traditional tools like NLTK primarily use word-level tokenisation (splitting text at whitespace and punctuation), BERT employs a more nuanced subword tokenisation strategy that offers several key advantages [49].

The tokeniser segments text into subword units, allowing it to handle out-of-vocabulary words effectively by breaking them into meaningful components. For example, ‘unaffordable’ might be split into ‘un’, ‘afford’, and ‘able’, preserving semantic meaning even for unfamiliar words. This approach proves particularly valuable for propaganda detection as it can capture nuanced word variations and neologisms often employed in persuasive text. Unlike NLTK’s more rigid tokenisation, BERT’s method learns from statistical patterns in training data, making it more adaptable to different writing styles and emerging language patterns.

Additionally, BERT’s tokeniser inherently captures morphological relationships between words, while NLTK requires separate stemming or lemmatisation steps to achieve similar results. That is why for this paper, steps such as whitespace removal and stemming were not performed when BERT’s tokeniser was used, but were for NLTK.

For the fragment level model, the NLTK tokeniser is used. NLTK was compared to BERT and did not improve results, so it was decided to use NLTK as it has significantly better computational efficiency.

3.3 Word Embeddings

Word embeddings map discrete words into continuous vector spaces, enabling mathematical computation of semantic relationships. While early approaches like Word2Vec and GloVe used static embeddings where each word had a fixed vector, BERT introduced contextual embeddings where a word's representation varies based on its surrounding context [49]. This allows BERT to capture nuanced meanings (like distinguishing between "dining table" and “data table") through its bidirectional transformer architecture. These embeddings can encode rich linguistic information including syntax and context-dependent semantics and therefore are often beneficial to model performance. While many related works utilised BERT embeddings with success, its implementation in the models presented here ultimately did not provide significant improvements. Use of embeddings also requires significantly more computational power, which is not conducive to a tool that needs to provide quick real-time analysis. For these reasons, word embeddings were ultimately not used in either model.

3.4 Model Architecture

3.4.1 CNN Implementation

The CNN implementation centers around a single 1D convolutional layer that processes embedded text sequences. The architecture starts with an embedding layer that maps the vocabulary tokens into dense 100-dimensional vectors. These embeddings feed into the core Conv1d layer which uses 100 output channels and a kernel size of 3, meaning it analyses trigrams (3-word windows) across the input sequence. The convolution operation slides this kernel across embedded text, creating feature maps that capture local patterns and relationships.

The model's forward pass transforms the embedded input to match CNN's expected input format. After applying convolution and ReLU activation, max pooling extracts the most salient features across the sequence length dimension. The resulting features then pass through a dropout layer (0.5 rate) for regularisation before the final linear classifier maps the 100-dimensional features to 2 output classes (propaganda and non-propaganda) for the propaganda detection task.

The CNN was only implemented for the first task, detection at the sentence level, as performance was overall worse than CRF so the CNN was not further developed for the second task.

3.4.2 CRF Implementation

The first CRF model for propaganda detection at the sentence level uses a linear-chain CRF with rich feature extraction including word position markers, capitalisation patterns, prefix/suffix information, and contextual windows. The model employs the LBFGS algorithm for optimisation with L1 and L2 regularisation (c1=0.1, c2=0.1). The feature set captures both local and sequential dependencies through 13 distinct feature functions including word identity, position flags, morphological features, and neighboring word context. It also employs early stopping based on validation performance by monitoring the F1-score on a held-out validation set to prevent overfitting. This method actively tracks the gap between training and validation performance, triggering early stopping when the validation performance stops improving for 3 epochs.

The second CRF implementation extends to multi-class sequence labeling using BIO (Beginning-Inside-Outside) tagging, incorporating an architecture that handles both span detection and technique classification. It maintains the same core feature extraction architecture but adapts to sequence labeling by tracking technique-specific transitions

using BIO tags, where ‘O’ tags mark non-propaganda while ‘B’ and ‘I’ tags denote the beginning and continuation of specific propaganda techniques.

The model identifies techniques through sequential labeling, leveraging the CRF’s ability to learn transition patterns between different technique tags and non-propaganda segments. The feature space remains consistent with the sentence level model but now operates on a more complex output space that captures both technique boundaries and classifications simultaneously. For example, it learns that an ‘I-technique’ tag must follow either a ‘B-technique’ tag of the same technique or another ‘I-technique’ tag, while a ‘B-technique’ tag can follow any other tag including ‘O’ tags for non-propaganda text. This design enables the model to learn intricate patterns of how different propaganda techniques manifest in text, including their boundaries, internal structure, and relationships with surrounding non-propaganda content. This implementation also utilises L1 and L2 regularisation as well as early stopping to prevent overfitting.

3.5 Training Methodology

3.5.1 Training Data Split

Data splitting is an important technique in ML where the available dataset is divided into separate subsets for training, validation, and testing. This allows for the ability to train a model on one portion of data, tune its hyperparameters using a validation set, and evaluate its performance on a completely unseen test set. The evaluation provides an unbiased estimate of the model’s generalisation capabilities and prevents overfitting. In both models, the initial dataset is split into these three sets using a two-step process. First, 80% of the data is allocated to the training set, then the remaining 20% is further split into validation and test sets again at an 80/20 ratio.

3.5.2 Cross-validation Approach

Cross-validation is a statistical technique to assess a model’s performance. It uses a process by which it divides the dataset into multiple subsets, trains the model on a portion of the data, and validates it on the remaining parts. This process is repeated several times with different data splits, providing a more robust evaluation of the model’s performance as compared to a single train-test split. Both models employ a cross-validation approach using a 5-fold strategy, by which the dataset is divided into 5 equal parts, the model is then trains on four parts and validated on the remaining part in each iteration. This process is repeated five times, allowing each fold to serve as the validation set once.

3.5.3 Early Stopping Mechanism

To prevent overfitting and optimise training efficiency, an early stopping mechanism is incorporated into the training process. During each iteration, the model's performance is evaluated on the validation set using F1-score. If no improvement in validation performance is observed for a specified number of epochs, in this case 3, training is halted. This approach helps prevent overfitting by ensuring that the model doesn't excessively adapt to the training data, and reduces the number of unnecessary computational costs by avoiding prolonged training when further improvements are unlikely.

3.5.4 Hyperparameter Configuration

Hyperparameter configuration involves selecting and setting optimal values for parameters that determine a model's behavior and performance. Unlike parameters learned from data during training, hyperparameters are predetermined and significantly influence the model's effectiveness. Both models employ the L-BFGS algorithm for optimisation, which is particularly effective for handling large-scale problems with many parameters.

L1 and L2 regularisation are techniques to prevent overfitting by adding penalty terms to the model's loss function. L1 regularisation pushes less important feature weights to zero, while L2 encourages all weights to remain small but non-zero. Higher values mean the model is pushed towards simpler solutions with smaller weights, whereas lower lets the model focus on fitting the training data more precisely. Both values are set to a moderate strength of 0.1 to get benefits of both approaches.

3.6 Evaluation Framework

3.6.1 Model Evaluation

All models were evaluated using sklearn's classification report which calculates precision, recall, accuracy, and F1-score for each class, as well as macro and weighted averages across all classes. The evaluation process examines both training and test performance to identify potential overfitting, specifically flagging cases where the different between training and test F1-scores exceed 0.1. The weighted averaging of metrics accounts for class imbalance in the dataset, ensuring that performance on less frequent classes contributes appropriately to the overall evaluation. The model's performance is

tracked during training using these metrics on the validation set, with the best performing model version being retained on validation F1-scores.

3.6.2 Comparative model analysis

To compare model performance, both the CNN and CRF were assessed based on classification report metrics. Optimal performance was deemed to be the highest combination of overall accuracy and consistency between classes. For example, if a model had high overall accuracy but a significant imbalance in accuracy between classes, this was considered worse performance compared to a model with lower overall accuracy but better balance between classes.

Since the goal of this task is accurate propaganda detection, it is important to achieve high precision to minimise false positives that could incorrectly label legitimate content as propaganda, while maintaining sufficient recall to effectively identify actual instances of propaganda. An imbalanced model that excels at identifying non-propaganda but frequently misses propaganda instances would fail to serve the fundamental purpose of the detection system, potentially allowing manipulative content to go unnoticed. This balanced approach ensures the system remains both reliable and trustworthy in real-world applications where false accusations of propaganda could be as problematic as failing to detect actual propaganda.

3.7 Israel-Palestine News Articles

3.7.1 Web Scraping Methodology

The article collection process employed a systematic web scraping methodology to gather news articles from the BBC and *The Guardian* websites across seven key topic areas related to the Israel-Palestine conflict. The approach targeted the topic-specific pages for Gaza, Israel, Palestine/Palestinian territories, West Bank, Hamas, Hezbollah, and Middle East news. The implementation incorporated randomised delays between requests and appropriate browser identification to minimise server load, ensuring robust and ethical data collection.

Approaches differed slightly between sources. The BBC scraper employed a dual-strategy approach, first attempting to fetch data from a structured data interface before falling back to webpage parsing. *The Guardian* scraper focuses on processing paginated topic sections and filtering out non-article content such as live blogs, videos, and

photo galleries. For both sources, the program extracted publication dates, headlines, URLs, and full article content. It also filtered for articles published after October 7, 2023. Duplicate detection was implemented to ensure articles were only collected once, even when appearing across multiple topic sections. The collected data from both sources was stored in a dataframe.

3.7.2 Data cleaning and preprocessing

The text cleaning process began with removing any residual HTML tags from the web scraping process using regular expressions. The program then implemented standardisation of typographical elements, converting various Unicode representations of quotation marks, dashes, and special characters to their standard ASCII equivalents. It also handled whitespace normalisation, removing redundant spaces and standardising line breaks. Duplicate lines within articles were identified and removed while preserving the narrative flow of the text.

The cleaned articles were then organised into a unified dataset, with each article assigned a unique identifier and labeled with its source publication. This process ensured that all text data maintained consistent formatting and character encoding while preserving the content and meaning of the original articles.

3.7.3 LLM Annotation

Due to the fact that the web scraping technique employed simply scraped all articles from the specified webpages after the specified date, a secondary step was performed to mark articles as relevant or irrelevant to the Israel-Palestine conflict. The classification program utilised modern language model embeddings using RoBERTa, employing a three-tiered keyword structure (see Figure 3), specifically tailored to identify content related to the Israel-Palestine conflict [50]. This approach leverages RoBERTa's contextual embeddings through mean pooling for semantic analysis of borderline cases and helps mitigate the limitations of simple keyword matching by incorporating semantic understanding through transformer-based embeddings. The program uses weighted thresholds for different keyword categories and combines these with cosine similarity scores against a reference text, allowing for nuanced classification decisions. It also uses a cascading decision structure, whereby articles with strong keyword evidence are immediately classified and those with ambiguous signals undergo additional semantic analysis. This methodology strikes a balance between computational efficiency and classification accuracy.

Articles were marked with ‘YES’ if the article was relevant to Israel-Palestine and ‘NO’ if they were irrelevant, allowing for specific analysis of relevant articles, as well as greater possibilities for analysis across these general topics.

```
TOPIC_KEYWORDS = {
    'primary': [
        'israel', 'palestine', 'gaza', 'hamas', 'idf', 'west bank',
        'jerusalem', 'tel aviv', 'palestinian', 'israeli'
    ],
    'secondary': [
        'conflict', 'war', 'attack', 'missile', 'rocket', 'ceasefire',
        'peace', 'negotiation', 'military', 'civilian', 'casualties',
        'humanitarian', 'settlement', 'border', 'security', 'middle east'
    ],
    'context': [
        'october 7', 'hostage', 'militant', 'refugee', 'aid',
        'violence', 'defense', 'prime minister', 'death toll',
        'diplomatic', 'un security council', 'resolution'
    ]
}
```

Fig 3.

4. Results

4.1 Initial Model Comparison - CNN vs CRF

Classification reports for the initial CNN and CRF models showed very similar results, with accuracy at 75% for both models. The CNN however, had a much greater class imbalance, with F1-score at 0.46 for the 'propaganda' class and 0.84 for 'non-propaganda'. The CRF on the other hand, had a more even split between two classes. For this reason, it was decided to continue using CRF for optimisation and further model development as it provided more balanced predictive performance across both classes, making it better suited for real-world propaganda detection where maintaining high sensitivity for both propaganda and non-propaganda content is crucial.

4.2 CRF Performance on Sentence-level classification

Initial results for sentence-level classification were very promising, with the model showing a weighted average of 0.85 across all metrics (see Table 1). After k-fold implementation, all metrics increase to 0.98. This was initially suspicious, as such a dramatic improvement warranted investigation. Further testing with 3 and 10 k-folds revealed that the model's performance was sensitive to fold size, returning to 0.85 with k=3 but maintaining 0.98 with k=10. The test set also performed higher than training by 0.3, which can be attributed to the upsampling strategy used during training, where the original distribution in the test set proved easier to classify than the artificially balanced training data. See confusion matrix in Figure 4.

Metric	Pre-kfold macro	Post-kfold macro	Pre-kfold weighted	Post-kfold weighted
Accuracy	0.85	0.98	0.85	0.98
Precision	0.86	0.98	0.85	0.98
Recall	0.84	0.98	0.85	0.98
F1	0.84	0.98	0.85	0.98

Table 1: Sentence classification before and after K-fold

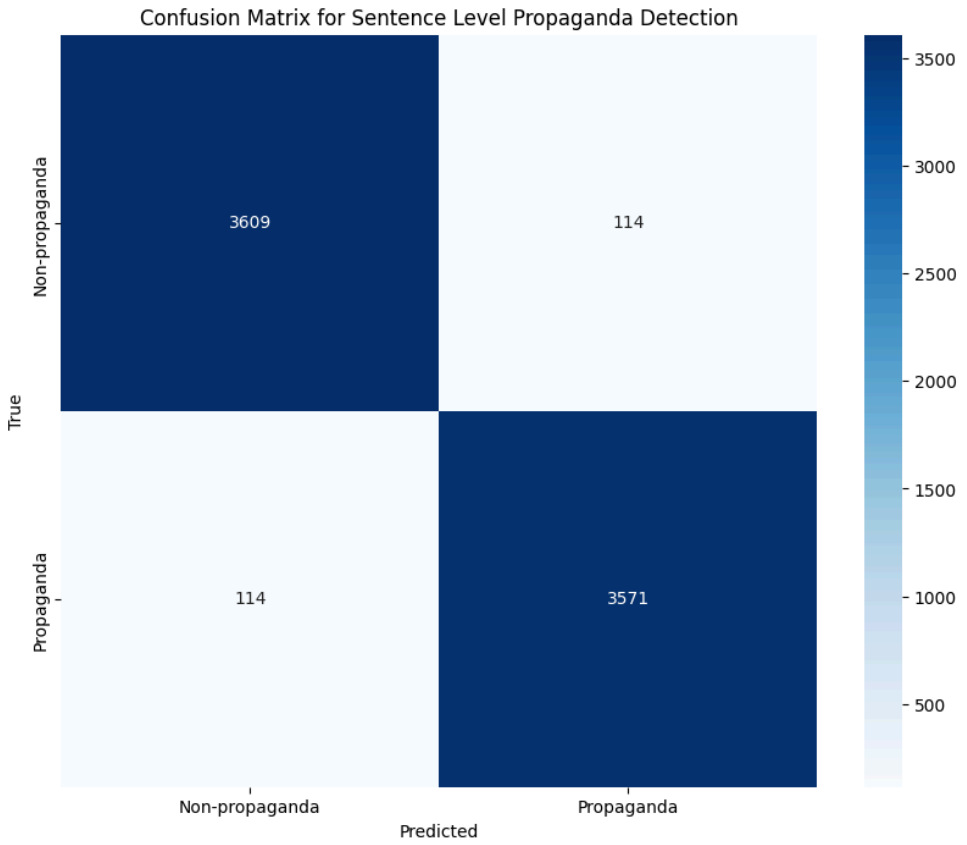

Fig. 4

4.3 CRF Performance on Fragment-level classification

Results for fragment-level extraction and classification were also significant, with the model showing a weighted average of 0.83 for accuracy, precision, and recall, and 0.82 for F1-score (see Table 2). F1-score across techniques was relatively consistent, with a minimum of 0.6 for *Loaded Language,* a maximum of 0.98 for *Thought-Terminating clichés*. K-fold implementation saw minimal improvement for accuracy and precision, with metrics increasing to 0.86. Recall and F1-score, however, saw a decrease to 0.61 and 0.71 respectively for macro average, but stayed higher and 0.86 and 0.85 for weighted average. See Table 3 for all technique scores.

This reduction suggests that the initial evaluation may have benefited from a particularly favorable train-test split, potentially overestimating the model's generalisation capabilities. Despite the lower recall, proceeding with the k-fold results is more methodologically sound because they provide a more realistic assessment of the model's performance across different data distributions. The higher accuracy and precision scores indicate that while the model makes fewer positive predictions overall, it maintains high

confidence in the predictions it does make. This trade-off between precision and recall is particularly acceptable in rhetorical technique classification, where false positives could be more problematic than false negatives in the downstream application of propaganda detection. Additionally, the k-fold results offer more robust evidence of the model's real-world applicability, as they better represent how it would perform across varying discourse styles and contexts.

Metric	Pre-kfold macro	Post-kfold macro	Pre-kfold weighted	Post-kfold weighted
Accuracy	0.83	0.86	0.86	0.98
Precision	0.83	0.86	0.86	0.98
Recall	0.83	0.61	0.86	0.98
F1	0.82	0.71	0.85	0.98

Table 2: Fragment classification before and after K-fold

Class	Precision	Recall	F1-score	Support
Appeal to Authority	0.94	0.62	0.75	4,696
Bandwagon, Reductio ad hitlerum	0.91	0.71	0.80	5,622
Causal_Oversimplification	0.92	0.66	0.77	6,118
Doubt	0.81	0.69	0.75	5,520
Exaggeration, Minimisation	0.77	0.51	0.61	2,439
Loaded Language	0.59	0.41	0.49	1,076
Name Calling, Labeling	0.75	0.35	0.48	1,467
O	0.85	0.98	0.91	95,099
Slogans	0.96	0.53	0.69	1,529
Flag-Waving	0.80	0.52	0.63	3,360
Whataboutism, straw man, red herring	0.99	0.58	0.73	5,657
Black-and-White Fallacy	0.97	0.58	0.73	4,867
Prejudice	0.80	0.77	0.78	5,166
Thought-terminating Cliches	0.99	0.64	0.78	2,324
accuracy			0.86	144,940
macro avg	0.86	0.61	0.71	144,940
weighted avg	0.86	0.86	0.85	144,940

Table 3: Classification Report for Fragment-level Test Set

4.4 Deployment on Israel-Palestine News Articles

The model analysed the provided news articles and provided a 'propaganda' or 'non-propaganda' label for each sentence in each article, as well as extracted propaganda sentence fragments and labelled them with a specified technique. See Figure 5 for sentence distribution.

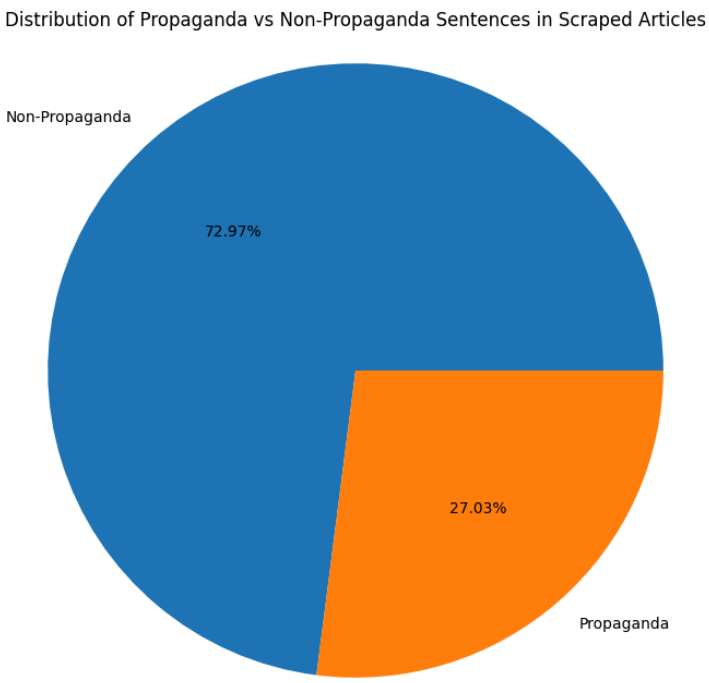

Fig. 5

Results showed that all articles on average contained 27.03% propaganda. This was calculated by taking sentences labeled as 'propaganda', dividing them by the total number of sentences per article, and averaging across all articles. Articles from the BBC had on average 26.87% of sentences with propaganda, vs *The Guardian's* 27.98%.

Looking at propaganda by sentence by topic, *The Guardian's* Middle East topic page had the highest propaganda score at 30.6% (see Table 4). No articles were collected from the BBC's Middle East page due to web scraping constraints. The next highest score is for Hamas at 28.33%, then Palestine at 28.13%, Gaza at 28.03%, West Bank at 26.25%, Israel at 26.2%, and Hezbollah at 24.5%. *The Guardian* scored higher for all topics besides Israel and Hezbollah, showing that there is a slightly higher chance one might encounter propaganda in articles from that source.

Topic	BBC	Guardian	Average
Middle East	NaN	30.60	30.60
Hamas	26.67	29.99	28.33
Palestine	27.91	28.34	28.13
Gaza	26.85	29.21	28.03
West Bank	25.51	26.99	26.25
Israel	27.39	25.02	26.20
Hezbollah	26.04	22.96	24.50

Table 4: Coverage Analysis of Middle East Topics in BBC and Guardian

Analysis of propaganda trends over time reveals relatively consistent levels of propaganda instances, averaging 27.26% of content across the period (see Figure 6). The data shows a few notable fluctuations, with a peak in May 2024 (29.04%) and low in August 2024 (23.31%). Though the volume of articles varies considerably throughout the period, from a high of 463 articles in October 2024 to a low of 76 articles in January 2024, averaging 196 articles per month. There is virtually no correlation (-0.033) between the number of articles and the percentage of propaganda content, suggesting that the intensity of propaganda remains consistent regardless of volume of coverage. The most stable period was from December 2023 to April 2024, where propaganda levels remained within a narrow band between 26.7% and 27.4%, despite fluctuating article counts. Note that this data was pulled on December 4th, 2024 so December statistics have been omitted from this analysis.

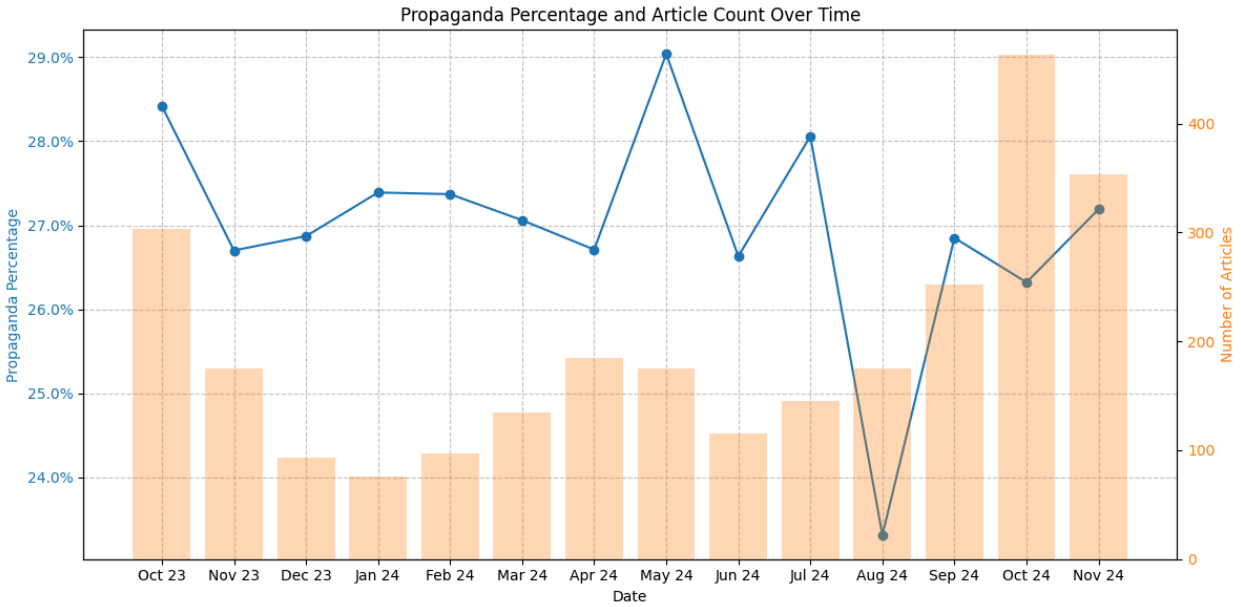

Fig. 6

Looking at propaganda techniques, 95.7% of sentences identified as propaganda did not have a technique associated (see table 5). There were cases like this in the training data, though a higher percentage would be expected considering that it is reasonable to presume that a propaganda sentence would employ at least one technique. This may be due to poor model performance on this unseen data, or that the technique used is not one captured in the 13 techniques. There were no instances of non-propaganda sentences also containing an identified span and technique.

Propaganda Sentences	Sentences with Technique	% of total
37,830	36,198	95.69%

Table 5: Propaganda Sentences vs Technique Spans

Loaded Language was by far the most identified technique, with it comprising 69.28% of all techniques identified (see Table 6). *Doubt* followed at 13%, then *Exaggeration/Minimisation* at 7.03%, *Name Calling/Labeling* at 3.6%, *Appeal to fear* at 3.35%, Flag waving at 1.53%, and all others below 1%. *The Guardian* uses more Loaded Language (71.96%) compared to the BBC (66.6%), while the BBC has higher instances of Doubt at 15.87% vs 10.28% for the BBC. The BBC uses *Exaggeration/Minimisation* more than *The Guardian* at 8.45% and 5.61%, respectively. Both publications make limited use of the less frequent techniques, which collectively account for less than 1% of all techniques employed. The most significant differences between the two sources show up in those top three techniques, with an average of 4.6% difference between the two.

Technique	BBC	Guardian	Average
Loaded Language	66.60	71.96	69.28
Doubt	15.87	10.28	13.08
Exaggeration, Minimisation	8.45	5.61	7.03
Name Calling, Labeling	2.98	4.21	3.60
Appeal to fear/prejudice	2.49	4.21	3.35
Flag-waving	1.66	1.40	1.53
Appeal to Authority	0.14	1.40	0.77
Causal Oversimplification	0.49	0.93	0.71
Black-and-White Fallacy	0.55	0.00	0.28
Bandwagon, reductio ad hitlerum	0.28	0.00	0.14
Whataboutism, straw man, red herring	0.28	0.00	0.14
Thought-terminating cliches	0.21	0.00	0.11

Table 6: Propaganda Techniques Analysis in BBC and Guardian Coverage

Examining techniques by topic, *Loaded Language* is again the most prevalent, ranging from 44.44% of all techniques in the Middle East topic page to 75.49% for West Bank (see Table 7). *Doubt* followed as the second most prevalent technique for all topic pages except Hezbollah, for which *Exaggeration/Minimisation* was second. West Bank, however, did not show any instances of *Doubt*. The most balanced distribution of techniques appears in the Israel topic page, which closely matches the average patterns, while the most divergent coverage appears in the Middle East and West Bank pages. Note again that there are no articles from the BBC's Middle East page which could contribute to this difference in pattern. Lower-frequency techniques are consistently rare across all topics.

Technique	Gaza	Hamas	Hezbollah	Israel	Middle East	Palestine	West Bank
Loaded Language	67.95	68.92	63.09	64.49	44.44	65.38	75.49
Doubt	14.09	16.04	12.08	13.08	33.33	16.78	16.67
Exaggeration, Minimisation	9.06	7.27	13.42	8.41	11.11	6.99	0.00
Name Calling, Labeling	2.68	2.76	4.70	5.61	5.56	3.50	0.98
Appeal to fear	2.68	2.26	2.01	0.93	0.00	3.85	4.90
Flag-waving	1.17	0.75	2.01	3.74	0.00	3.15	0.98
Appeal to Authority	0.50	0.00	0.00	0.93	5.56	0.00	0.00
Causal Oversimplification	0.50	0.50	1.34	0.93	0.00	0.00	0.98
Black-and-White Fallacy	0.67	0.50	1.34	0.00	0.00	0.00	0.00
Bandwagon, reductio ad hitlerum	0.34	0.50	0.00	0.00	0.00	0.00	0.00
Whataboutism, straw man, red herring	0.34	0.50	0.00	0.00	0.00	0.00	0.00
Thought-terminating cliche	0.00	0.00	0.00	1.87	0.00	0.35	0.00

Table 7: Propaganda Techniques Analysis by Topic in Middle East Coverage

For articles that were labelled as relevant to Israel-Palestine, 26.92% of sentences were marked as propaganda. Irrelevant articles were at 24.86%, showing only a 2% difference between the two. This may lead us to believe that articles pertaining to Israel-Palestine do not have a significantly higher instance of propaganda compared to all articles on the specified topics.

5. Conclusion

5.1 Discussion

5.1.2 Final CRF Model for Propaganda Detection

Results show that a CRF model like the one developed in this paper has the potential to be effectively deployed for real-word propaganda detection in news articles. The model showed significant performance on both the sentence and fragment-level tasks, though the fact that over 95% of propaganda sentences did not have an associated fragment indicate further refinement is necessary for fragment extraction and identification. Additionally, while high performance is the desired outcome, such high levels of accuracy are also cause for suspicion. In the results for the 2020 Sem-Eval11 task, the highest F1-score for the test set was 51.54%, precision was 66.88%, and recall at 80.37%. The results of this model are more consistent with these precision and recall scores for the fragment-level task, but significantly higher for F1-score. This discrepancy indicates that while the model seems to excel at identifying specific instances of propaganda when found, it may be overfitting to certain patterns or missing more subtle variations of propagandistic content.

Positively, the low computational cost and fast processing time indicate that this model could be suitable for a real-time analysis tool that could be used by the general public to easily analyse news articles while reading them. This would provide them with instant feedback they can flag potentially biased content and critically evaluate the content they're consuming. Though further refinement and testing against more new, unseen data is certainly necessary before deploying it as a public-facing tool.

5.1.3 Israel-Palestine News Analysis

The consistent level of propaganda techniques across different sources, topics, and time periods, suggests a potentially systemic feature of contemporary news coverage rather than intentional bias. Especially considering the lack of correlation between article volume and propaganda percentage, this consistency raises questions about whether these rhetorical devices are inherent to modern journalistic practice rather than deliberate propaganda efforts.

Regarding specific techniques, while *The Guardian* shows a higher propensity for Loaded Language and the BBC more towards Doubt, the overall similarity on their

propaganda percentages suggest that these represent different writing styles rather than varying levels of bias.

The finding that Israel-Palestine relevant articles only show marginally higher propaganda levels compared to irrelevant articles challenges assumptions about heightened rhetorical manipulation in news coverage. The consistency across sources and topics, combined with the relatively modest impact of conflict-related content on propaganda levels, indicates that these techniques might be better understood as features of modern journalistic language rather than deliberate attempts at manipulation. However, this interpretation warrants closer examination, as the articles were drawn from general Middle East/Palestine/Israel coverage. This is a topic area that historically attracts heightened propaganda techniques due to the interplay between Western foreign policy interests and domestic social attitudes towards different ethnic and religious groups in the region. This context raises important questions about how the baseline for 'neutral' reporting might differ across various geographic and cultural contexts in Western media coverage.

5.2 Drawbacks

While the CRF model provided a workable foundation, its sequential nature may have limited the capture of long-range dependencies and contextual relationships that are important for propaganda identification. The model's reliance on local features potentially missed broader structures and narrative patterns.

The dataset, while comprehensive, was still relatively small compared to what would be ideal for a task of this complexity. In the actual SemEval-2020 Task 11 there was an additional development set that teams deployed their models on, which they then submitted for scoring. Only the scorers had access to the true labels for this data, which were not included in the GitHub repository where this data was gathered. This lack of data from what was originally supplied means that this model worked off an even smaller dataset than those related works. Additionally, the decision was made to remove the technique Repetition as it required additional programming techniques that were outside the scope of this paper. This could have contributed to the lack of propaganda sentences with an identified technique if that technique was Repetition.

A significant limitation arose from computational constraints that prevented the full exploration of transformer-based architectures. While initial experiments with BERT word embeddings didn't show promising results, the computational demands proved prohibitive for further testing and optimisation. This restricted the ability to leverage such language

models that have demonstrated superior performance in similar tasks. The inability to test multiple iterations or alternative word embeddings may have prevented the discovery of more optimal representations for propaganda detection. Additionally, more extensive hyperparameter tuning such as grid searches were impacted, potentially resulting in suboptimal model configurations.

5.3 Ethical Considerations

While valuable for promoting media literacy, automated propaganda detection tools could be misused for censorship or to unfairly discredit legitimate journalism. The binary classification of content as “propaganda” or “non-propaganda” may oversimplify the complex nature of news reporting, where perspectives and interpretations can vary significantly. Additionally, the model’s high accuracy rates require careful scrutiny to ensure they don’t result from overfitting or biases in the training data.

The deployment of this system to analyse Israel-Palestine coverage raises additional ethical considerations. While the model wasn’t trained on this specific topic, applying automated detection tools to such sensitive topics requires careful consideration of context and potential impacts. Using coverage from specific outlets like the BBC and *The Guardian* as test cases could provide valuable insights, but the findings must be presented with appropriate caveats about the limitations of automated analysis. Questions of transparency and accountability are also important, as users of such a detection tool should understand their limitations and potential biases, rather than treating the model’s outputs as definitive judgements about journalistic integrity.

5.4 Future Work

In future iterations of this model, it may be interesting to combine CNN and CRF architectures. Integrating a CNN for feature extraction with a CRF for sequence labeling could leverage the CNN’s ability to capture local patterns and hierarchical features with the CRF’s strength in modeling dependencies between adjacent labels. It would also be beneficial to revisit BERT embeddings and investigate why they yielded worse results, as well as try out other embeddings such as GloVe or word2vec. Additionally, as with any ML model, training the model on more data would likely improve its performance. This is especially true given the complex and nuanced nature of propaganda detection where exposure to a wider variety of examples is paramount in ensuring accurate generalisation to new data.

With regards to analysis of Israel-Palestine news coverage, it would be beneficial to compare the results of the articles pulled for this analysis vs articles on different topics or from different parts of the world. As mentioned, this is a topic area that historically attracts heightened propaganda techniques. So while propaganda in this data was consistent in almost every aspect, the overall percentage in this topic and geographical space could be higher compared to coverage of less contentious issues or regions. Future research could explore whether certain propaganda techniques are more prevalent in different types of coverage, and whether the intensity or frequency of propaganda varies systematically across different subject matters. It would also be interesting to compare coverage before and after October 7, 2023, to see if the initiation of this new phase of conflict altered baseline propaganda levels in news coverage of the conflict or region.

A major goal would be to integrate this model into a real-world tool, such as a browser extension, that would allow users to get instant and visual feedback on the online news they are consuming. This tool could scan a webpage and extract the relevant content, then run the text through the model and provide an interactive display whereby users can see which sentences were identified as propaganda, and which specific fragments along with their techniques were identified. The tool could also provide an overall propaganda likelihood percentage to give readers an estimate on the amount of propaganda they are consuming, while presenting these insights in a gentle, exploratory manner rather than as definitive judgements. This approach acknowledges that we all have inherent biases and that questioning our deeply-held beliefs is a gradual process. It encourages users to view the tool as a companion in critical thinking rather than an authority challenging their perspectives. By framing the detection of propaganda techniques as an opportunity for reflection rather than a criticism, users may feel more comfortable engaging with information that challenges their existing viewpoints.

References

1. E. Aïmeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: A review," *Social Network Analysis and Mining*, vol. **13**, pp. 30, 2023.
2. BBC Monitoring, "Israel-Palestinian conflict: False and misleading claims fact-checked," BBC News, May 15, 2021. [Online].
3. A. Ray and J. F. George, "Online disinformation and the psychological bases of prejudice and political conservatism," in *Proc. 52nd Hawaii Int. Conf. System Sciences*, 2019, pp. 2742–2752.
4. H. Rashkin, E. F. Bell, Y. Choi, and S. Volkova, "Multilingual connotation frames: A case study on social media for targeted sentiment analysis and forecast," in *Proc. ACL*, 2017, pp. 2073.
5. A. Barrón-Cedeño, I. Jaradat, G. Da San Martino, and P. Nakov, "Proppy: Organizing the news based on their propagandistic content," *Information Processing & Management*, vol. **56**, no. 5, pp. 1849–1864, 2019.
6. G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, and P. Nakov, "SemEval-2020 task 11: Detection of propaganda techniques in news articles," in *Proc. 14th Workshop on Semantic Evaluation*, 2020, pp. 1377–1391.
7. Y. Can, "Digital deception: Disinformation's impact in the Israel-Hamas war," Wilson Center, Oct. 20, 2023. [Online].
8. A. M. Guess and B. A. Lyons, "Misinformation, disinformation, and online propaganda," in *The Oxford Handbook of Political Communication*, K. Kenski and K. H. Jamieson, Eds. Cambridge University Press, 2020, pp. 379–399.
9. B. L. Smith, "Propaganda: Definition, history, techniques, examples, & facts," Encyclopedia Britannica, Aug. 27, 2024. [Online].
10. J. A. Tucker, A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan, "Social media, political polarization, and political disinformation: A review of the scientific literature," Hewlett Foundation, Mar. 2018.
11. S. Zannettou, M. Sirivianos, J. Blackburn, and N. Kourtellis, "The web of false information," *Journal of Data and Information Quality*, vol. **11**, no. 3, pp. 1–37, 2019.
12. T. R. Glander, "Origins of mass communications research during the American Cold War: Educational effects and contemporary implications," 1999.
13. H. Ingram, "A brief history of propaganda during conflict: A lesson for counter-terrorism strategic communications," *Terrorism and Counter-Terrorism Studies*, 2016.
14. B. N. Porter, *Trees, Kings, and Politics*. Saint-Paul, 2003.
15. J. Grethlein, *The Greeks and Their Past*. Cambridge University Press, 2010.

16. P. Zanker, *The Power of Images in the Age of Augustus*, A. Shapiro, Trans. University of Michigan Press, 1988.
17. M. Camille, *The Gothic Idol: Ideology and Image-Making in Medieval Art*. Cambridge University Press, 1989.
18. M. U. Edwards, *Printing, Propaganda, and Martin Luther*. Fortress Press, 2005.
19. P. Feuerherd, "The uncommon sense of Thomas Paine," JSTOR Daily, Jan. 10, 2017. [Online].
20. H. D. Lasswell, *Propaganda Technique in The World War*. Knopf, 1927.
21. Z. A. B. Zeman, *Nazi Propaganda*, 2nd ed. London, 1973.
22. N. J. Cull, V. Gatov, P. Pomerantsev, A. Applebaum, and A. Shawcross, "Soviet subversion, disinformation and propaganda: How the West fought against it," LSE Institute of Global Affairs, 2017.
23. S. Woolley and P. N. Howard, *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford University Press, 2019.
24. A. Vindman, "The Russian disinformation threat: Active campaigns in 2024," Kettering Foundation, Oct. 10, 2024. [Online].
25. M. H. Elmasry, "Images of the Israel-Gaza war on Instagram: A content analysis of Western broadcast news posts," *Journalism & Mass Communication Quarterly*, 2024.
26. American Jewish Committee, "A guide to recognizing when anti-Israel actions become antisemitic." [Online].
27. EU DisinfoLab, "The role of media in producing and spreading disinformation campaigns." [Online].
28. M. Smith, "Which media outlets do Britons trust in 2023?" YouGov, May 25, 2023. [Online].
29. E. Fondren, "We are propagandists for democracy: The Institute for Propaganda Analysis' pioneering media literacy efforts to fight disinformation (1937–1942)," *American Journalism*, vol. **38**, no. 3, pp. 258–291, 2021.
30. O. Gavrilenko, Y. Oliinyk, and H. Khanko, "Analysis of propaganda elements detecting algorithms in text data," Springer International Publishing, 2020.
31. G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov, "Fine-grained analysis of propaganda in news articles," arXiv:1910.02517, 2019.
32. RAND Corporation, "Tools that fight disinformation online," 2022. [Online].
33. IBM, "What is machine learning?" 2024. [Online].
34. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
35. T. Mitchell, *Machine Learning*. McGraw-Hill, 1997.
36. P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. **55**, no. 10, pp. 78–87, 2012.

37. C. Sutton, "An introduction to conditional random fields," *Foundations and Trends in Machine Learning*, vol. **4**, no. 4, pp. 267–373, 2012.
38. S. Sun et al., "A survey of optimization methods from a machine learning perspective," arXiv:1906.06821, 2019.
39. M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. **45**, no. 4, pp. 427–437, 2009.
40. D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Dorling Kindersley, 2014.
41. D. G. Jones, "Detecting propaganda in news articles using large language models," *Eng OA*, vol. **2**, no. 1, pp. 01–12, 2024.
42. P. SuthanthiraDevi, S. Karthika, K. Sowmya, S. Srinidhi, and S. Pavithra, "Detection of propaganda from news articles using deep learning," *Int. J. Advanced Trends in Computer Science and Engineering*, vol. **9**, no. 3, pp. 3500–3505, 2020.
43. M. Salman, A. Hanif, S. Shehata, and P. Nakov, "Detecting propaganda techniques in code-switched social media text," in *Proc. EMNLP*, 2023, pp. 1044.
44. A. M. U. D. Khanday, M. A. Wani, S. T. Rabani, Q. R. Khan, and A. A. Abd El-Latif, "HAPI: An efficient hybrid feature engineering-based approach for propaganda identification in social media," *PLOS ONE*, vol. **19**, no. 7, 2024.
45. J. Piskorski, N. Stefanovitch, N. Nikolaidis, G. Da, and P. Nakov, "Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques," in *Proc. ACL*, 2023, pp. 169.
46. D. Dimitrov et al., "Detecting propaganda techniques in memes," in *Proc. 59th Annual Meeting of the Association for Computational Linguistics*, 2021, pp. 6603–6617.
47. Media Bias Fact Check, "Media bias/fact check," 2024. [Online].
48. A Data Pro, "Quality information insight to power strategic, tactical and smart decisions," 2022. [Online].
49. P. De Camillis, "Analysing natural language processing techniques: A comparative study of NLTK, spaCy, BERT, and DistilBERT on customer query datasets," M.S. thesis, CCT College Dublin, 2022.
50. Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, 2019.

Appendix

Propaganda Techniques

1. **Loaded language**: Using specific words and phrases with strong emotional implications (either positive or negative) to influence an audience.
2. **Name calling or labeling**: Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable, or loves and praises.
3. **Exaggeration or minimisation**: Representing something in an excessive manner (making things larger, better, worse) or making something seem less important or smaller than it actually is.
4. **Doubt**: Questioning the credibility of someone or something.
5. **Appeal to fear/prejudice**: Building support for an idea by instilling anxiety and/or panic in the population towards an alternative, often based on preconceived judgements.
6. **Flag-waving**: Playing on strong national or group feelings (e.g., race, gender, political preference) to justify or promote an action or idea.
7. **Causal oversimplification**: Assuming a single cause or reason for an issue when there are multiple causes. This includes scapegoating, transferring blame to one person or group without investigating complexities.
8. **Slogans**: A brief and striking phrase that may include labeling and stereotyping, often serving as an emotional appeal.
9. **Appeal to authority**: Claiming a statement is true simply because a valid authority or expert on the issue supports it, without any other supporting evidence.
10. **Black-and-white fallacy**: Presenting two alternative options as the only possibilities, ignoring other potential solutions.
11. **Thought-terminating cliché**: Words or phrases that discourage critical thought and meaningful discussion on a topic, offering seemingly simple answers to complex questions or distracting attention.
12. **Whataboutism**: Discrediting an opponent's position by charging them with hypocrisy without disproving their argument.
    Straw man: Substituting an opponent's proposition with a similar one that is easier to refute.
    Red herring: Introducing irrelevant material to divert attention from the main issue.

13. **Bandwagon**: Persuading the audience to join a cause because “everyone else is doing it.”
    Reductio ad hitlerum: Discrediting an idea or action by associating it with groups or figures that are despised by the target audience.